



Human-Centered AI Transparency: Bridging the Socio-Technical Gap

Q. Vera Liao

Microsoft Research
University of Michigan



Microsoft



UNIVERSITY OF
MICHIGAN

ChatGPT's inaccuracies are causing real harm

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users.
That's No Laughing Matter

AI Bias In Recruitment: Ethical Implications And Transparency

AP

AI chatbots are supposed to improve health care. But research says some are perpetuating racism

Snapchat's AI chatbot may pose privacy risk to children, says UK watchdog

Ethical concerns mount as AI takes bigger decision-making role in more industries

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison

WHO warns against bias, misinformation in using AI in healthcare

How Artificial Intelligence Can Deepen Racial and Economic Inequities

Here are dozens of ways AI could be used for harm — and some too scary to test

NEWSLETTERS • EYE ON A.I.

ChatGPT's inaccuracies are causing real harm

Ethical concerns mount as AI takes bigger decision-making role in more industries

TECH • ARTIFICIAL INTELLIGENCE

The New AI-Powered Bing Is Threatening Users.
That's No Laughing Matter

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You
to Prison

Responsible AI (RAI): Develop and deploy AI technologies responsibly by mitigating harms

RAI requires approaching AI as sociotechnical systems

Inequities

Snapchat's AI chatbot may pose privacy risk to children, says UK watchdog

Here are dozens of ways AI could be used for harm — and some too scary to test

The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility

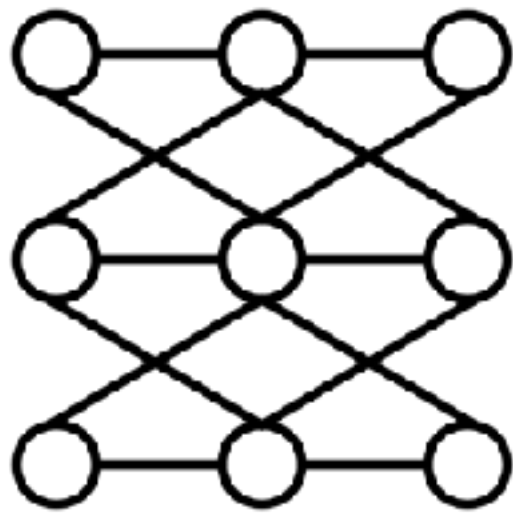
Mark S. Ackerman
University of California, Irvine

ABSTRACT

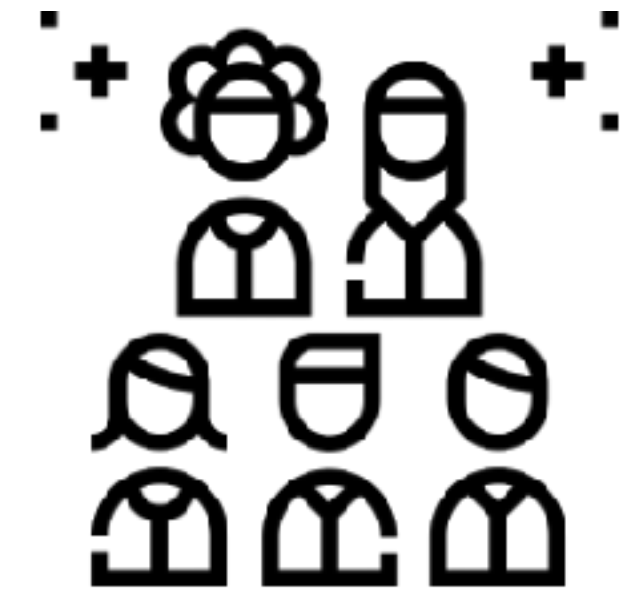
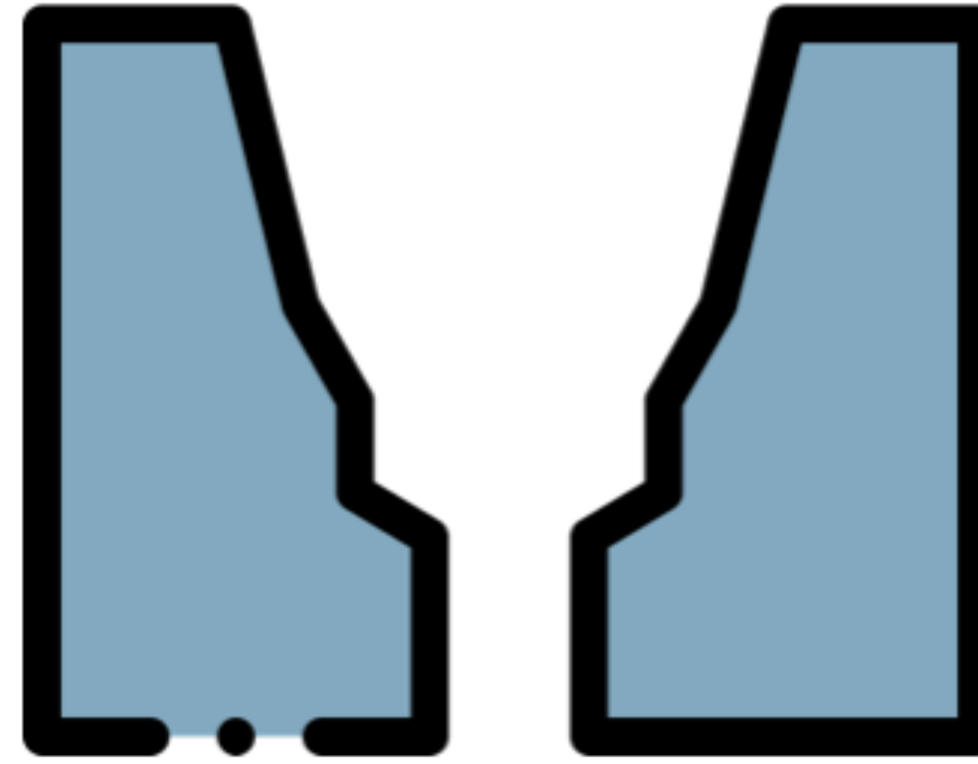
Over the last 10 years, Computer-Supported Cooperative Work (CSCW) has identified a base set of findings. These findings are taken almost as assumptions within the field. In summary, they argue that human activity is highly flexible, nuanced, and contextualized and that computational entities such as information sharing, roles, and social norms need to be similarly flexible, nuanced, and contextualized. However, current systems cannot fully support the social world uncovered by these findings. In this article I argue that there is an inherent gap between the social requirements of CSCW and its technical mechanisms. The *social-technical gap* is the divide between what we know we must support socially and what we can support technically. Exploring, understanding, and hopefully ameliorating this social-technical gap is the central challenge for CSCW as a field and one of the central problems for human-computer interaction. Indeed, merely attesting the continued centrality of this gap could be one of the important intellectual contributions of CSCW. I also argue that the challenge of the social-technical gap creates an opportunity to refocus CSCW.

Socio-technical gap:
Inevitable gap between what computational systems can do and what people need in social contexts

Socio-Technical Gap

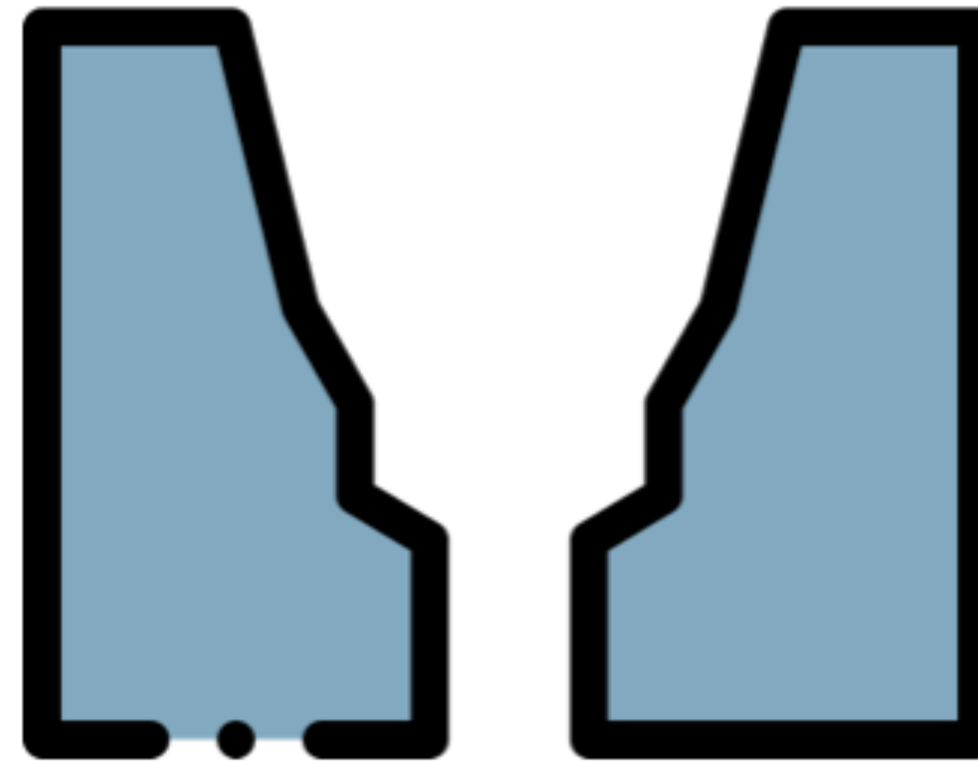


Precise, formal, abstract...

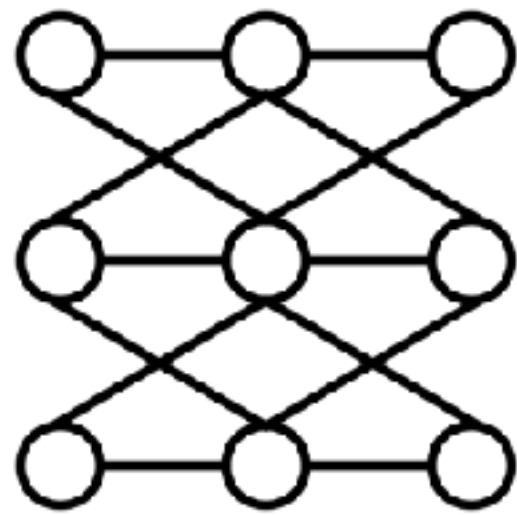


Contextual, nuanced, fluid...

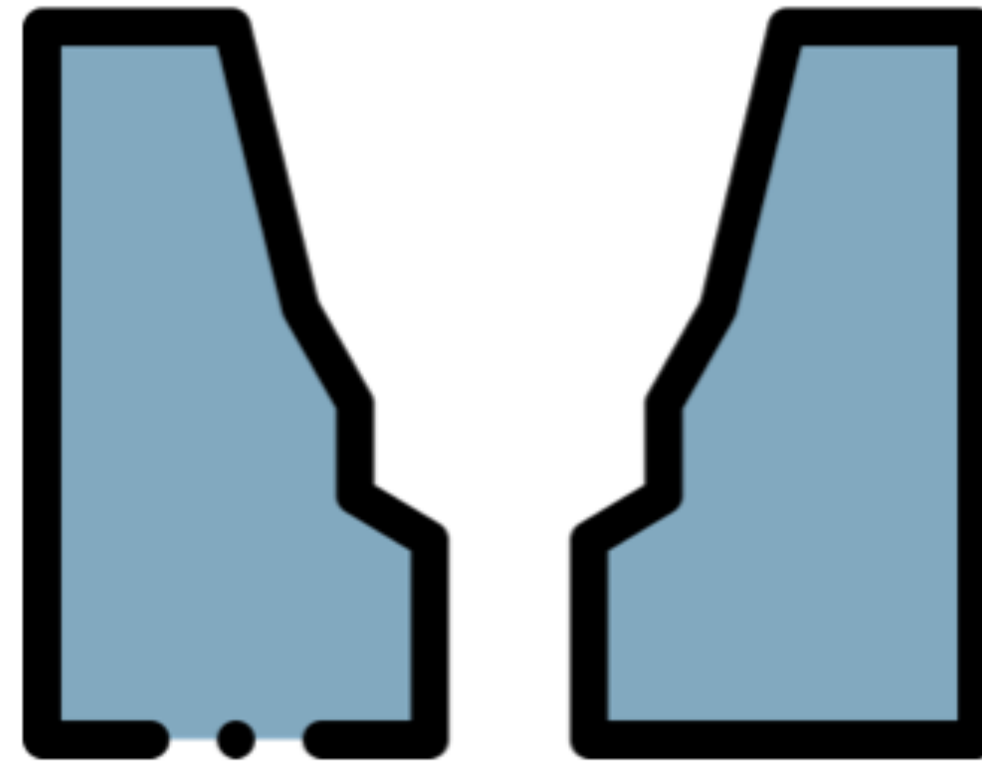
Socio-Technical Gap in the Age of AI



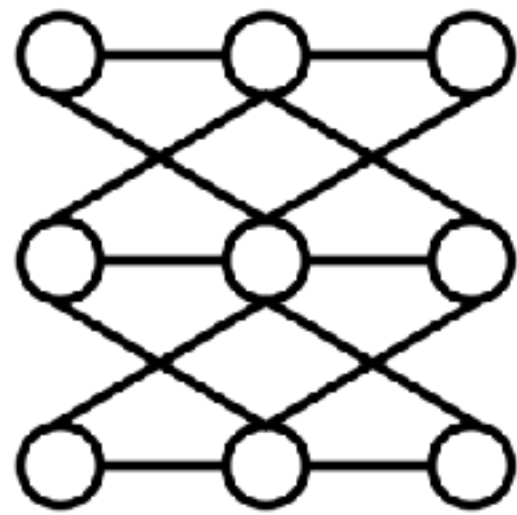
Socio-Technical Gap in the Age of AI



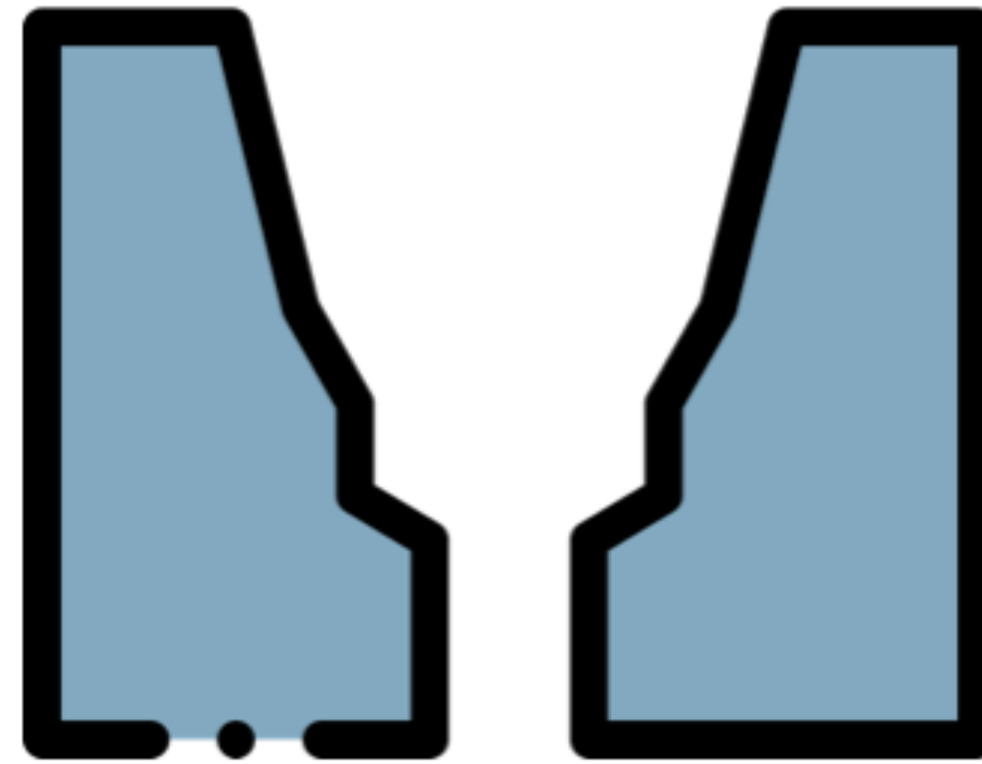
AI techniques
developed in research



Socio-Technical Gap in the Age of AI



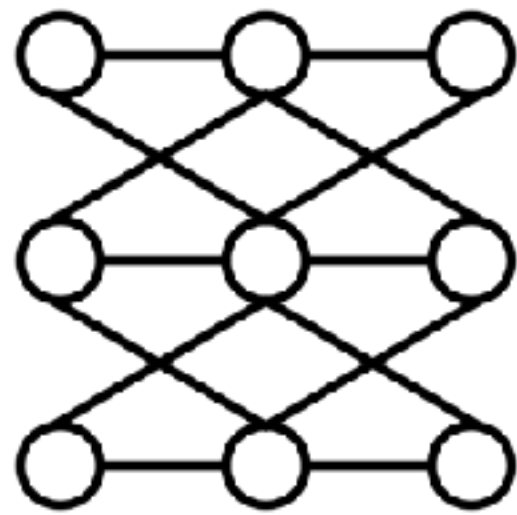
AI techniques
developed in research



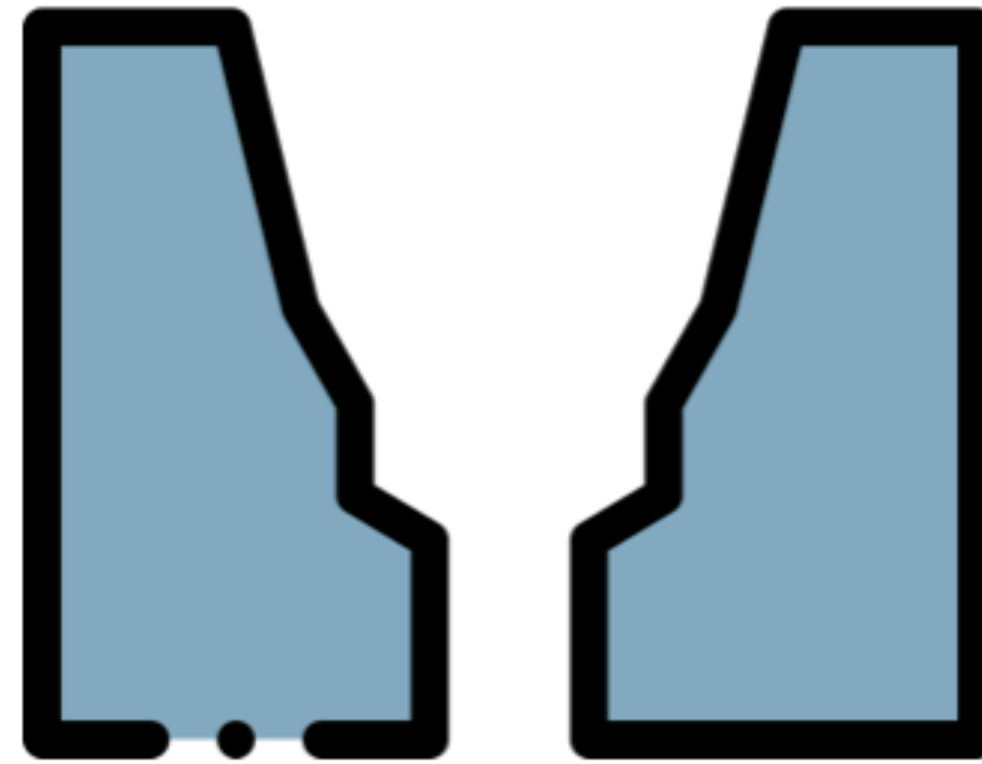
Target use case

Failure in accounting for **human interaction**

Socio-Technical Gap in the Age of AI



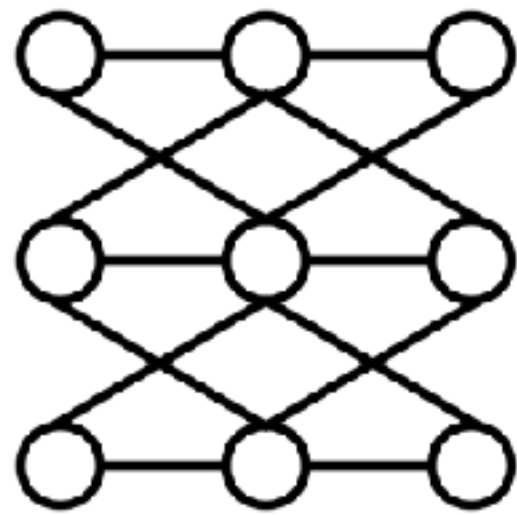
AI techniques
developed in research



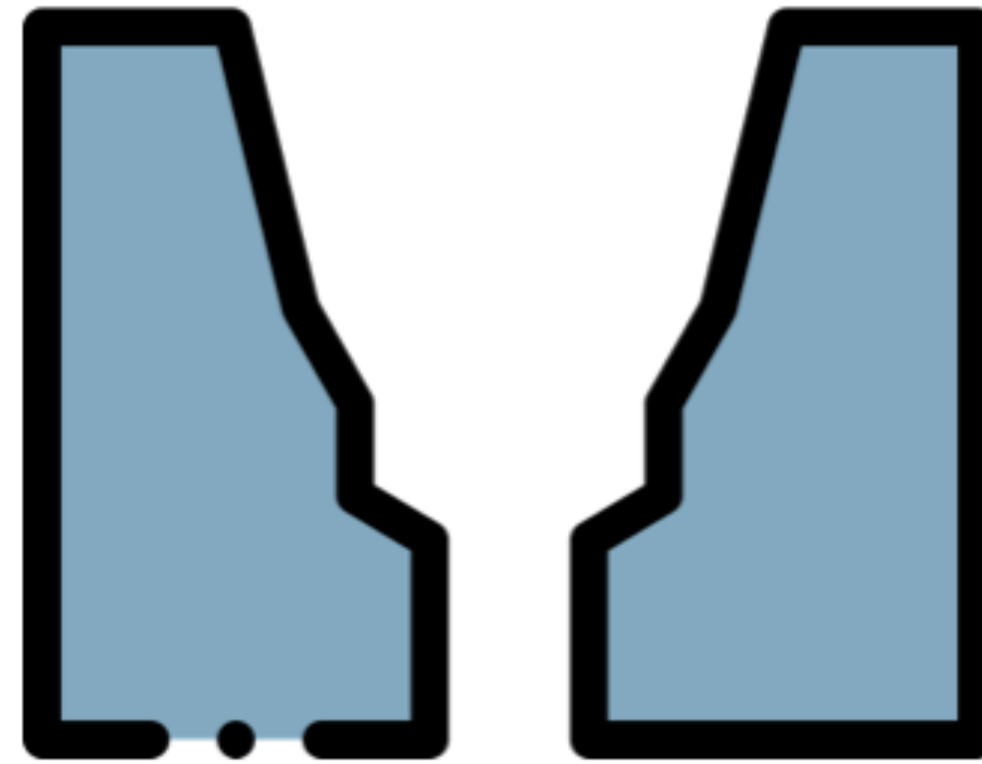
Heterogeneous applications
in diverse contexts

Failure in anticipating **diverse and new use cases**

Socio-Technical Gap in the Age of AI



Precise, formal, abstract...



Contextual, nuanced, fluid...



Failure in formalizing “socio” requirements

(Context-specific human requirements for a “good” technology)



The socio-technical gap of AI is widening in the age of “large” models, as the underlying AI technologies become increasingly homogeneous

Failures contributing to socio-technical gap of AI

Failure in accounting for
human interaction

Failure in anticipating new
and diverse use cases

Failure in formalizing “socio”
requirements

Failures contributing to socio-technical gap of AI

Failure in accounting for human interaction

Failure in anticipating new and diverse use cases

Failure in formalizing “socio” requirements

How can HCI research help to address these failures?

Empirical evaluation

Formative research

Theory and framework building

A Crash Course on **AI Transparency**

From black box to white box: Reclaiming human power in AI

We Need AI That Is Explainable, Auditable, and Transparent

Why transparency is key to unlocking AI's full potential

How the EU AI Act Can Increase Transparency Around AI Training Data

Companies Grapple With AI's Opaque Decision-Making Process

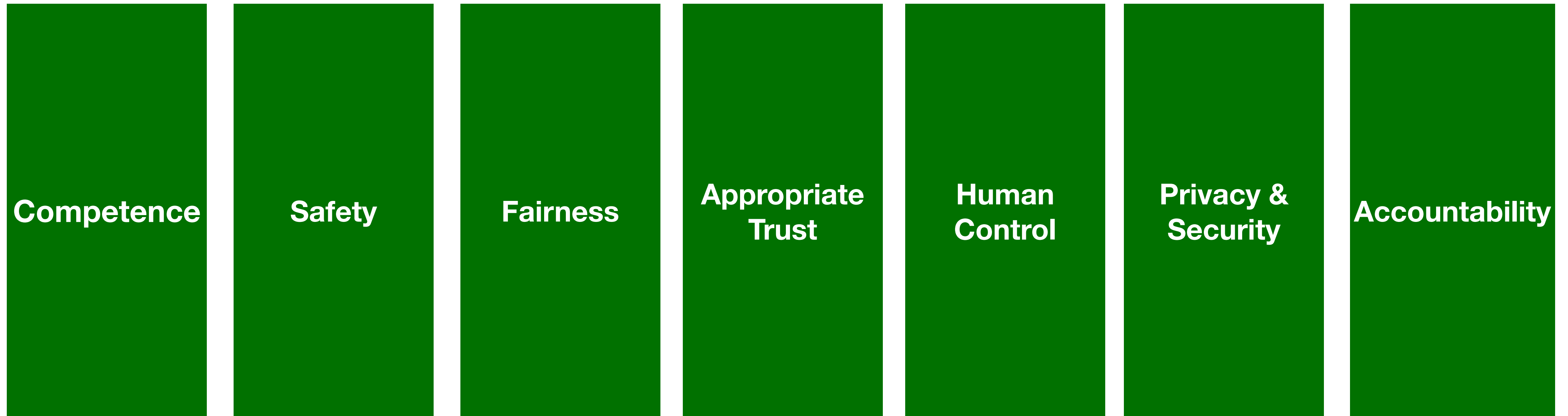
**NEW CALIFORNIA LAW WILL REQUIRE AI
TRANSPARENCY AND DISCLOSURE MEASURES**

Transparency: Disclosing information about a technology to enable **human understanding**

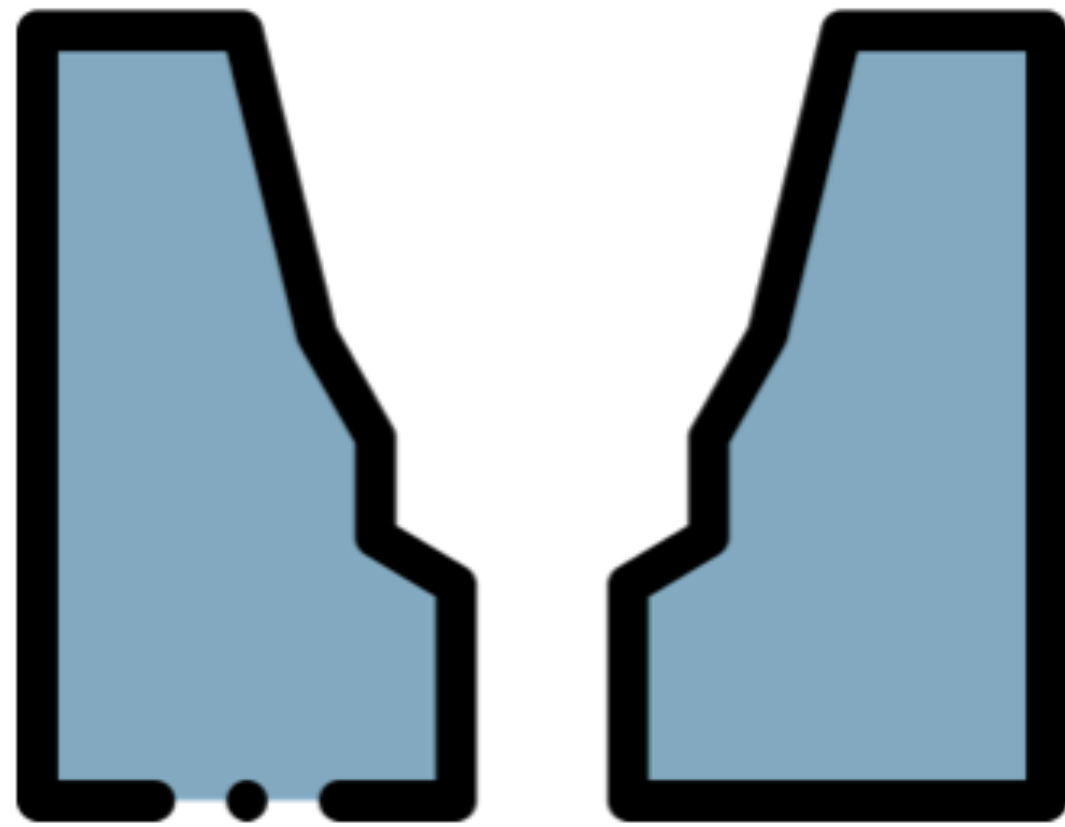


Microsoft Responsible AI Principles

Understanding as Means to Many Ends of RAI



Understanding AI allows for human control, effective interaction and oversight



Socio-technical gap of AI
transparency: **how
might well-intentioned
transparency
approaches go wrong?**

Four transparency approaches:

- **Model reporting**
- **Evaluation**
- **Explainability**
- **Uncertainty**

Four transparency approaches:

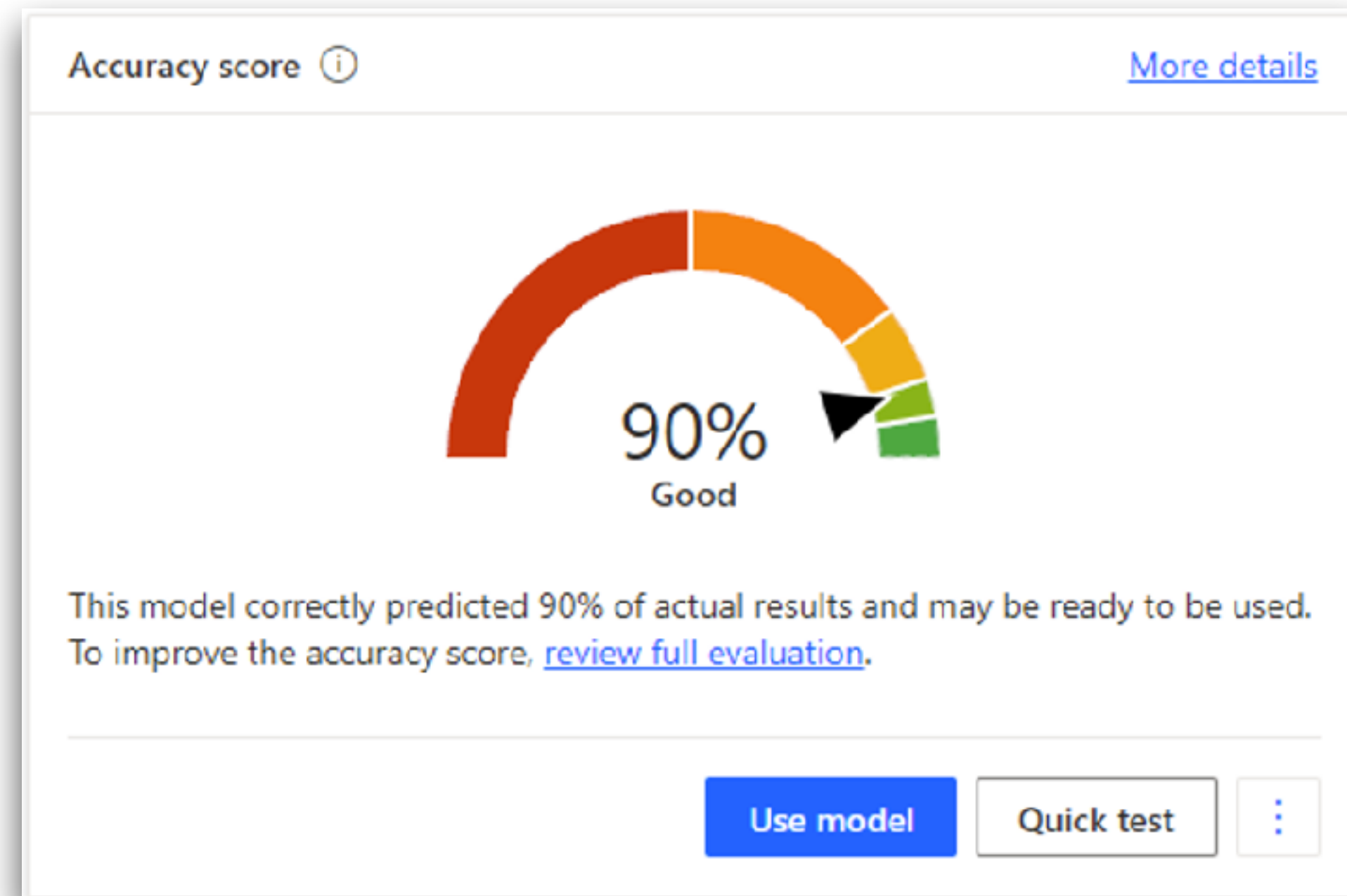
- **Model reporting**
- Evaluation
- Explainability
- Uncertainty

Model Card
• Model Details. Basic information about the model. – Person or organization developing model – Model date – Model version – Model type – Information about training algorithms, parameters, fairness constraints or other applied approaches, and features – Paper or other resource for more information – Citation details – License – Where to send questions or comments about the model • Intended Use. Use cases that were envisioned during development. – Primary intended uses – Primary intended users – Out-of-scope use cases • Factors. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3. – Relevant factors – Evaluation factors • Metrics. Metrics should be chosen to reflect potential real-world impacts of the model. – Model performance measures – Decision thresholds – Variation approaches • Evaluation Data. Details on the dataset(s) used for the quantitative analyses in the card. – Datasets – Motivation – Preprocessing • Training Data. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets. • Quantitative Analyses – Unitary results – Intersectional results • Ethical Considerations • Caveats and Recommendations

Model Card

Four transparency approaches:

- Model reporting
- **Evaluation**
- Explainability
- Uncertainty

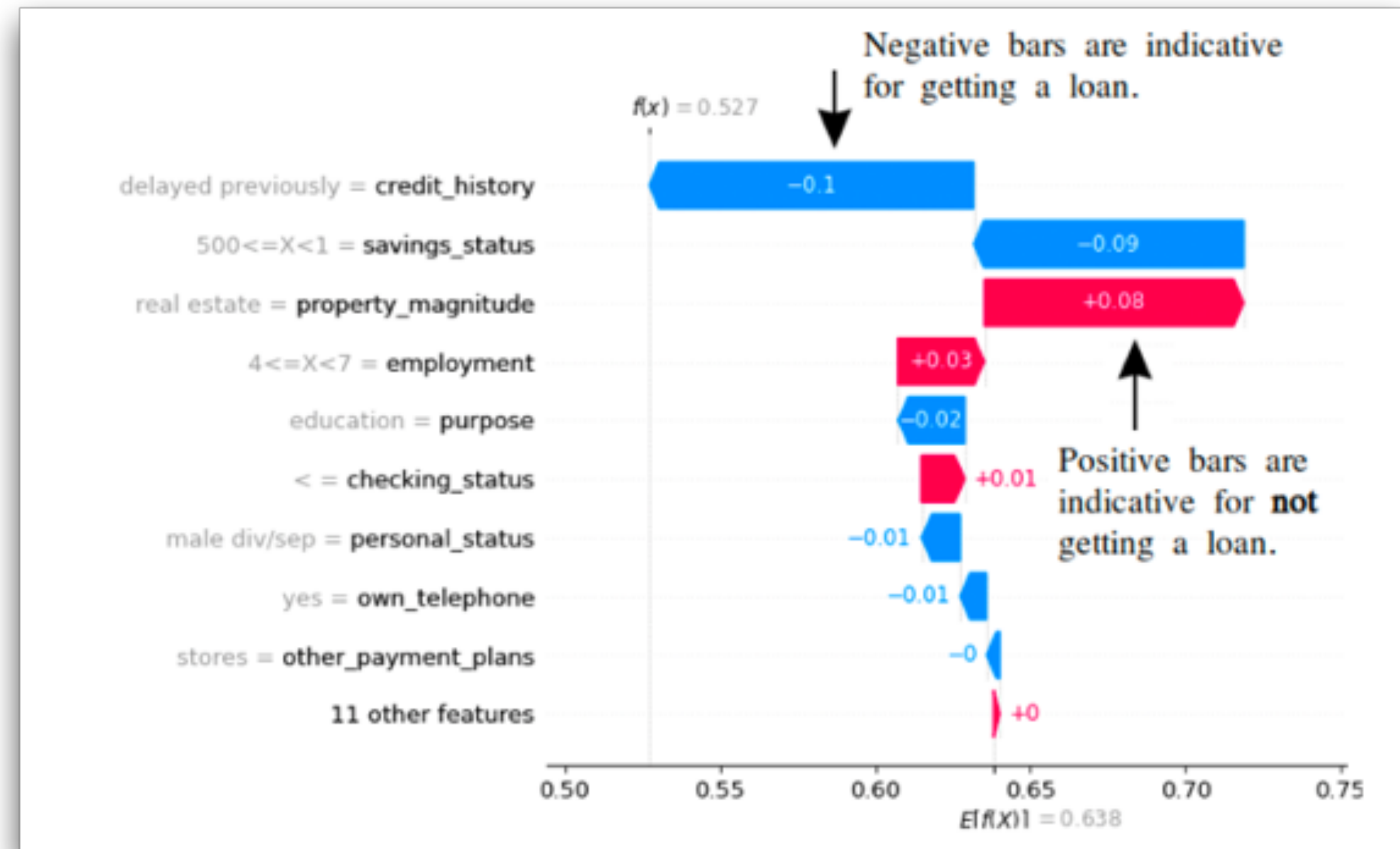


Accuracy Score

Four transparency approaches:

- Model reporting
- Evaluation
- **Explainability**
- Uncertainty

Prediction: **Low risk**



Feature Importance Explanation

For a loan risk assessment tool
(Hadash et al., CHI 2022)

Four transparency approaches:

- Model reporting
- Evaluation
- **Explainability**
- Uncertainty

_ is currently vice president of the American Society of Retina Specialists. Dr. _ receives grant support for multicenter clinical trials sponsored by the National Institutes of Health, Genentech and Regeneron.

AI Prediction: professor

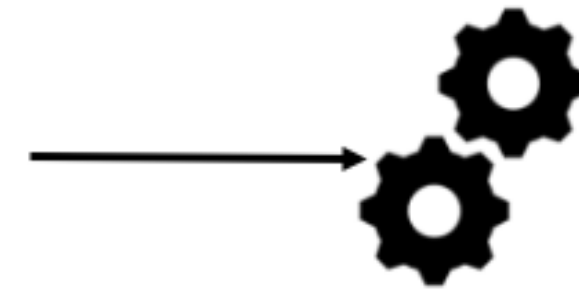
Feature Importance Explanation

For bio classification
(Chen et al., CSCW 2023)

Four transparency approaches:

- Model reporting
- Evaluation
- Explainability
- **Uncertainty**

Customer: Jason
Assets score: **88**
No. Of satisfactory trades: **0**
Mo. since account open: **3**
Number of inquiries: **1**
Debt percentage: **10%**



**Risk of failing
to repay: low**

Confidence: 60%

Confidence Score

**Failures contributing to
socio-technical gap of AI**

**How can HCI research help
address these failures?**

**Failure in accounting for
human interaction**

Empirical evaluation

Failure in anticipating new
and diverse use cases

Formative research

Failure in formalizing “socio”
requirements

Theory and framework
building

What Can Empirical Evaluation Do?



Identify limitations and pitfalls,
and understand *why* to inform
better approaches

Compare technical paradigms to
guide research efforts

Challenge the underlying
assumptions about use



Transparency
= Providing Information
= Understanding?

Assuming “perfect”
information processing

Assuming “rational” use
of information

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model.

how much the human understands a model’s behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Abstract

Explainable AI (XAI) Technique: Feature Importance Explanation Stated Goal: Appropriate Reliance

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron^{1,2} Hugo Touvron^{1,3} Ishan Misra¹ Hervé Jegou¹
Julien Mairal² Piotr Bojanowski¹ Armand Joulin¹

¹ Facebook AI Research ² Inria* ³ Sorbonne University

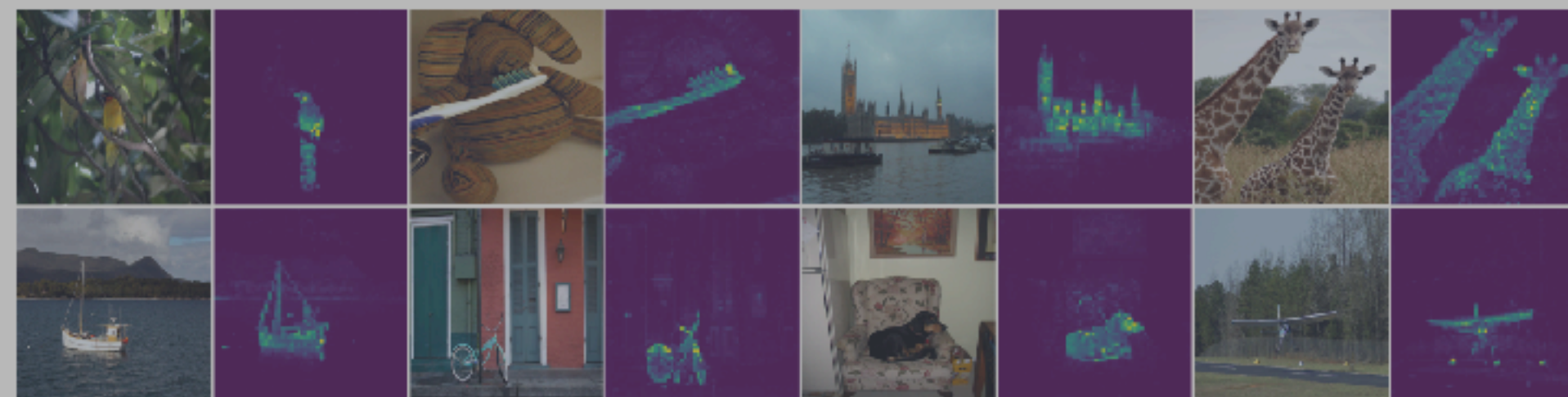


Figure 1: Self-attention from a Vision Transformer with 8×8 patches trained with no supervision. We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju^{1*} Michael Cogswell¹ Abhishek Das¹ Ramakrishna Vedantam^{1*}
Devi Parikh^{1,2} Dhruv Batra^{1,2}

¹Georgia Institute of Technology ²Facebook AI Research
{ramprs, cogswell, abhshkdz, vrma, parikh, dbatra}@gatech.edu

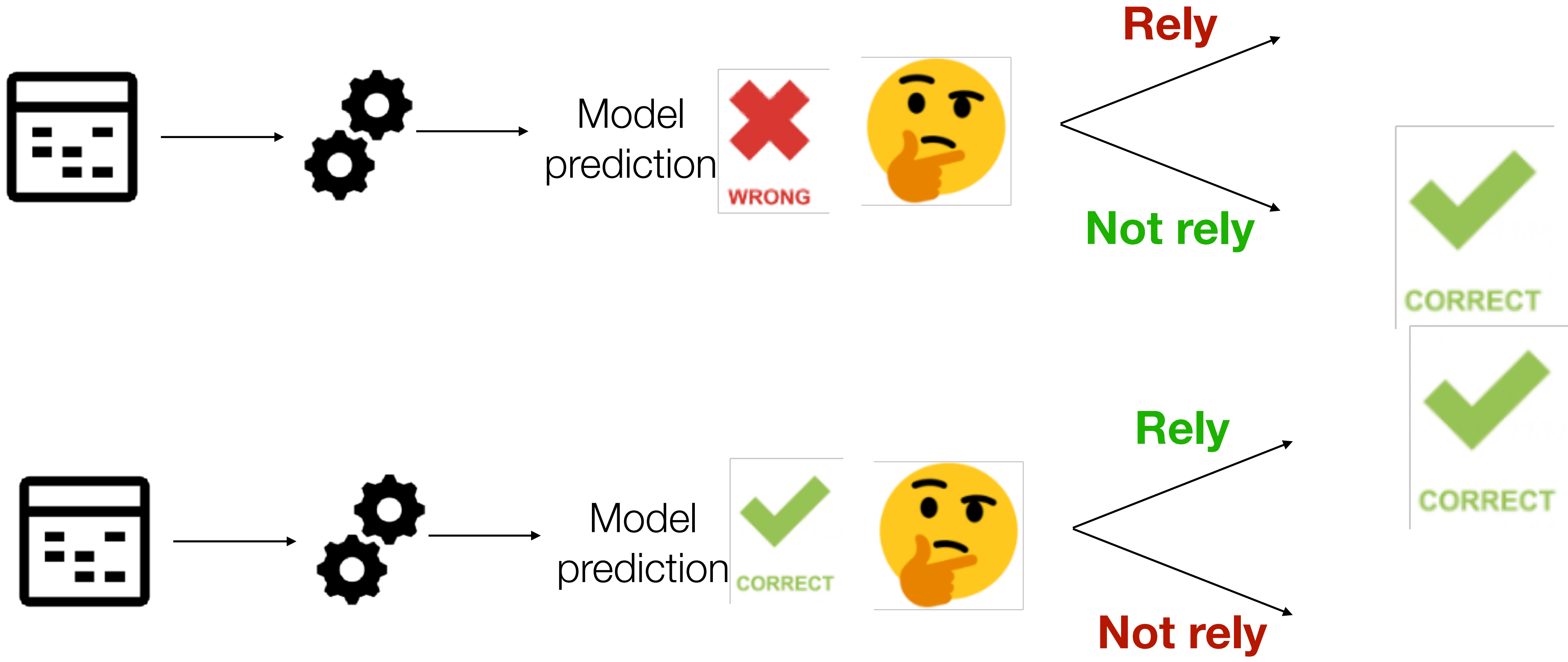
Abstract

We propose a technique for producing ‘visual explanations’ for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say logits for ‘dog’ or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families:

1. Introduction

Convolutional Neural Networks (CNNs) and other deep networks have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification [24, 16] to object detection [15], semantic segmentation [27], image captioning [43, 6, 12, 21], and more recently, visual question answering [3, 14, 32, 36]. While these deep neural networks enable superior performance, their lack of decomposability into *intuitive and understandable* components makes them hard to interpret [26]. Consequently, when

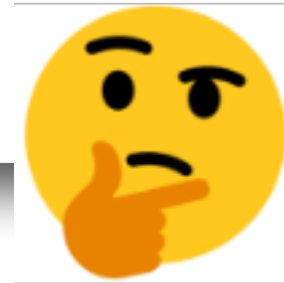
Appropriate Reliance is Key to AI-Assisted Decision Making



Explainable AI Pitfall: Overreliance on AI

_ is currently vice president of the American Society of Retina Specialists. Dr. _ receives grant support for multicenter clinical trials sponsored by the National Institutes of Health, Genentech and Regeneron.

AI Prediction: professor

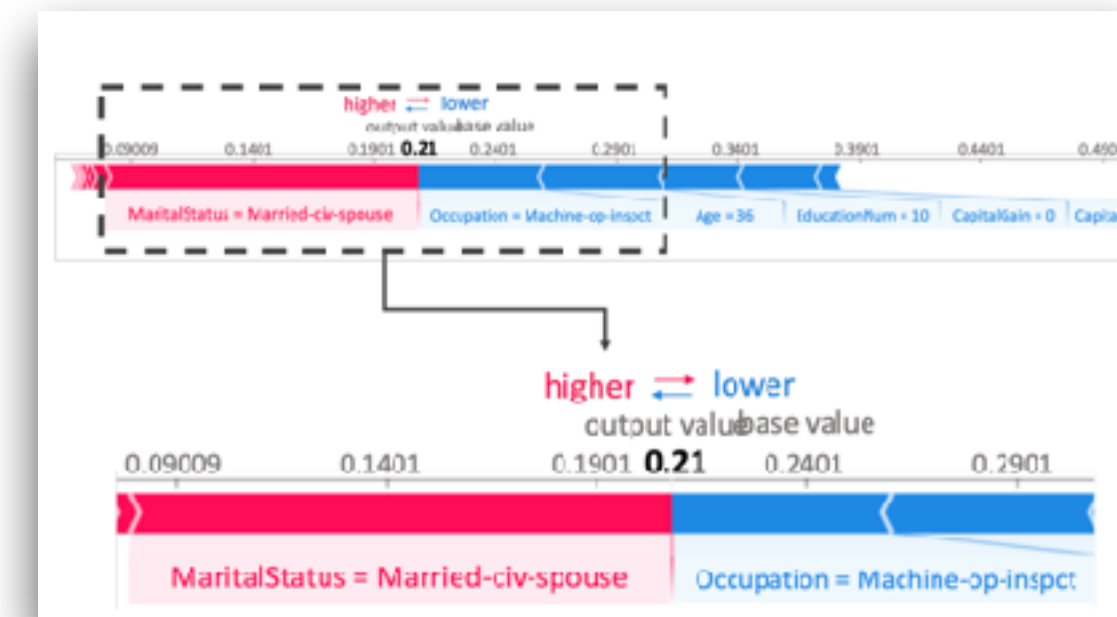
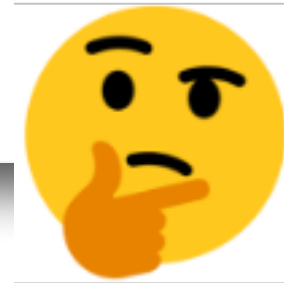


Explanations increase
overreliance for AI-assisted
decision making and reduce
decision performance

Explainable AI Pitfall: Overreliance on AI

_ is currently vice president of the American Society of Retina Specialists. Dr. _ receives grant support for multicenter clinical trials sponsored by the National Institutes of Health, Genentech and Regeneron.

AI Prediction: professor



Explanations increase overreliance for AI-assisted decision making and reduce decision performance

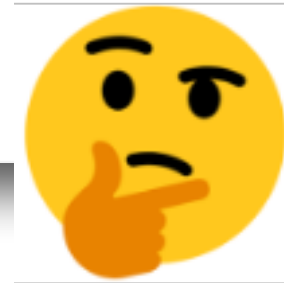
XAI debugging tools lead to over-confidence in model readiness for deployment

Explainable AI Pitfall: Overreliance on AI

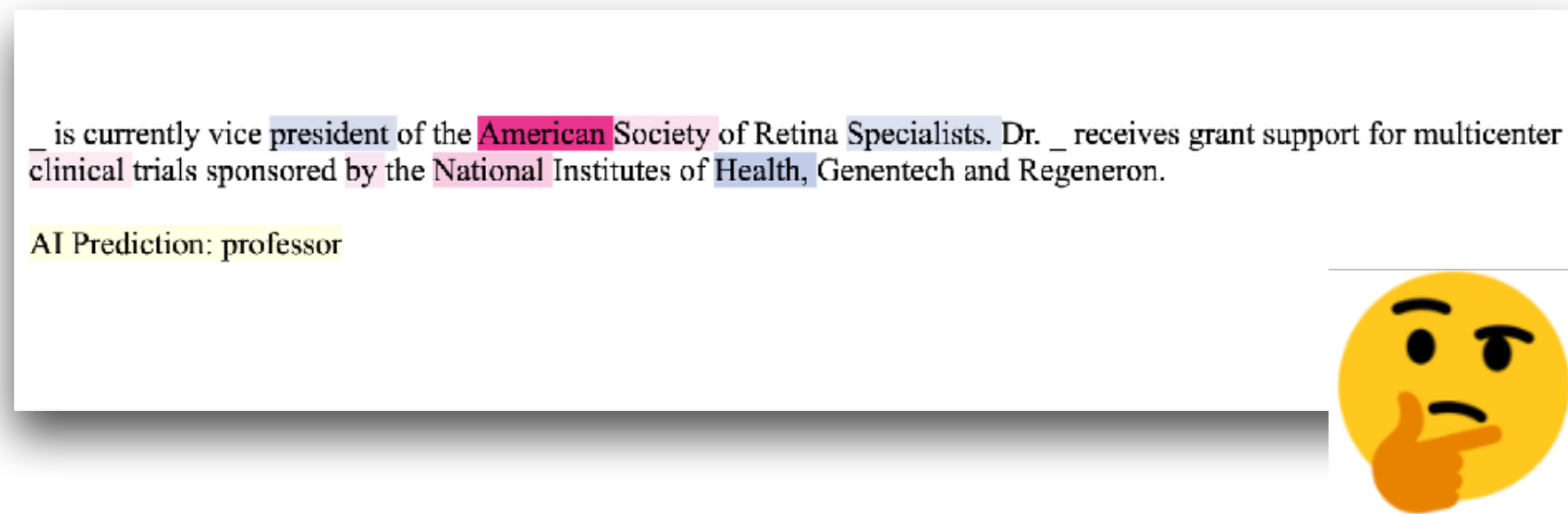
People may not be able to or motivated to engage, but invoke heuristics (*explainable=capable*)

_ is currently vice president of the American Society of Retina Specialists. Dr. _ receives grant support for multicenter clinical trials sponsored by the National Institutes of Health, Genentech and Regeneron.

AI Prediction: professor



Explainable AI Pitfall: Overreliance on AI



People may not be able to or motivated to engage, but invoke heuristics (*explainable=capable*)

When they engage, explanation is overwhelming and disruptive of their thinking process

Precise quantification of features is not how people reason

PROTOTYPE SELECTION FOR INTERPRETABLE CLASSIFICATION

BY JACOB BIEN¹ AND ROBERT TIBSHIRANI²

Stanford University

This paper is dedicated to the memory of Sam Roweis

Prototype methods seek a minimal subset of samples that can serve as a distillation or condensed view of a data set. As the size of modern data sets grows, being able to present a domain specialist with a short list of “representative” samples chosen from the data

Examples are not Enough, Learn to Criticize!
Criticism for Interpretability

Been Kim*
Allen Institute for AI
beenkim@csail.mit.edu

Rajiv Khanna
UT Austin
rajivak@utexas.edu

Oluwasanmi Koyejo
UTUC
sanmi@illinois.edu

Abstract

Example-based explanations are widely used in the effort to improve the interpretability of highly complex distributions. However, prototypes alone are rarely

Example/prototype-based explanation: a less “popular” area of XAI

The Bayesian Case Model: A Generative Approach
for Case-Based Reasoning and Prototype
Classification

Been Kim, Cynthia Rudin and Julie Shah
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
{beenkim, rudin, julie_a_shah}@csail.mit.edu

Abstract

We present the Bayesian Case Model (BCM), a general framework for Bayesian case-based reasoning (CBR) and prototype classification and clustering. BCM brings the intuitive power of CBR to a Bayesian generative framework. The BCM learns not only the “unimportant” observations that best represent classes in

This Looks Like That: Deep Learning for
Interpretable Image Recognition

Chaofan Chen*
Duke University
cfchen@cs.duke.edu

Oscar Li*
Duke University
oscarli@alumni.duke.edu

Chaofan Tao
Duke University
chaofan.tao@duke.edu

Alina Jade Barnett
Duke University
abarnett@cs.duke.edu

Jonathan Su
MIT Lincoln Laboratory†
su@ll.mit.edu

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

Abstract

When we are faced with challenging image classification tasks, we often explain our reasoning by dissecting the image, and pointing out prototypical aspects of one class or another. The mounting evidence for each of the classes helps us make our final decision. In this work, we introduce a deep network architecture – *prototypical part network* (ProtoPNet), that reasons in a similar way: the network dissects the image by finding prototypical parts, and combines evidence from the

Similar Bio #1

_ received _ B.A. degree from the University of the State of New York, _ M.A. from the University of Kent at Canterbury, and _ Ph.D. in Political Science from Cambridge University. _ has been a Fulbright Fellow to Korea and Indonesia.

AI Prediction: professor

Actual Profession: professor

Similar Bio #2

_ areas of expertise include adolescent and young adult education, teacher leadership, experiential and adventure education, and professional development. _ received _ B.S. from Kent State University (1999) before earning _ M.Ed. (2004) from Ashland University and _ Ph.D. (2007) from Kent State University. Before joining Wooster's faculty in 2007, _ was an adjunct faculty member for The College of Wooster and Kent State University, served as a K-12 curriculum and instruction specialist, and middle school science, social studies, and language arts teacher. _ has written and presented on numerous topics in _ fields of expertise.

AI Prediction: professor

Actual Profession: professor

Example-Based Explanation

Example-Based Explanation is a More Human-Compatible Paradigm

Similar Bio #1

_ received _ B.A. degree from the University of the State of New York, _ M.A. from the University of Kent at Canterbury, and _ Ph.D. in Political Science from Cambridge University. _ has been a Fulbright Fellow to Korea and Indonesia.

AI Prediction: professor

Actual Profession: professor

Similar Bio #2

_ areas of expertise include adolescent and young adult education, teacher leadership, experiential and adventure education, and professional development. _ received _ B.S. from Kent State University (1999) before earning _ M.Ed. (2004) from Ashland University and _ Ph.D. (2007) from Kent State University. Before joining Wooster's faculty in 2007, _ was an adjunct faculty member for The College of Wooster and Kent State University, served as a K-12 curriculum and instruction specialist, and middle school science, social studies, and language arts teacher. _ has written and presented on numerous topics in _ fields of expertise.

AI Prediction: professor

Actual Profession: professor

Flexibility of not being disrupted

Support inductive reasoning to verify or discover new features

Salient uncertainty signals

Example-Based Explanation

Reducing Conversational Agents’ Overconfidence Through
Linguistic Calibration

Sabrina J. Mielke^{1,2} Arthur Szlam² Emily Dinan² Y-Lan Boureau²

¹Department of Computer Science, Johns Hopkins University, USA ²Facebook AI Research, USA
sjmielke@jhu.edu {aszlam, edinan, ylan}@fb.com

Abstract

While improving neural dialogue agents’ factual accuracy is the object of much research,

control of the alignment between the confidence of an answer and its likelihood of being correct seems like a promising type of remediation: It

Uncertainty in Natural Language Generation:
From Theory to Applications

Joris Baan^{1*} Nico Daheim^{2*} Evgenia Ilia^{1*} Dennis Ulmer^{3,4*} Haau-Sing Li²
Raquel Fernández¹ Barbara Plank^{9,3} Rico Sennrich^{7,8} Chrysoula Zerva^{5,6} Wilker Aziz¹

¹University of Amsterdam ²TU Darmstadt & hessian.AI ³IT University of Copenhagen
⁴Pioneer Centre for Artificial Intelligence ⁵Instituto de Telecomunicações
⁶Instituto Superior Técnico & LUMIS (Lisbon ELLIS Unit) ⁷University of Zurich
⁸University of Edinburgh ⁹LMU Munich & Munich Center for Machine Learning

Abstract

Recent advances of powerful Language Models have allowed Natural Language Generation

Touvron et al., 2023a,b). Their widespread use makes it increasingly important to build NLG systems that are trustworthy and representative of the diversity of its users (Bhatt et al., 2021; Joshi et al., 2022).

Uncertainty Technique: Uncertainty Expression Generation
Stated Goal: Appropriate Reliance

Teaching models to express their uncertainty in words

Stephanie Lin
University of Oxford

sylin07@gmail.com

Jacob Hilton
OpenAI

jhilton@openai.com

Owain Evans
University of Oxford

owaine@gmail.com

Abstract

We show that a GPT-3 model can learn to express uncertainty about its own answers in natural language without use of model logits. When given a question, the model generates both an answer and a level of confidence (e.g. “90% confidence” or “high confidence”). These

SEMANTIC UNCERTAINTY: LINGUISTIC INVARIANCES
FOR UNCERTAINTY ESTIMATION IN NATURAL
LANGUAGE GENERATION

Lorenz Kuhn, Yarin Gal, Sebastian Farquhar
OATML Group, Department of Computer Science, University of Oxford
lorenz.kuhn@cs.ox.ac.uk

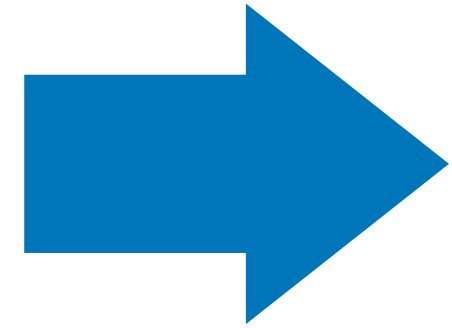
ABSTRACT

We introduce a method to measure uncertainty in large language models. For tasks like question answering, it is essential to know when we can trust the natural language outputs of foundation models. We show that measuring uncertainty in natural language is challenging because of ‘semantic equivalence’—different sentences can mean the same thing. To overcome these challenges we introduce

Natural-Language Uncertainty Expression

Confidence: 60%

Numerical Expression



Question: What is the capital of Mauritania?

LM Expressions of Confidence

Plain Statement

Ø

It's Nouakchott.

Weakener

I'm not sure, maybe

it's Nouakchott.

Verbal Expression

Q. Is Spironolactone an FDA-approved drug for treating acne?

Not uncertain

Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].

Uncertain in the
first-person perspective

I'm not sure, but my guess is *Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].*

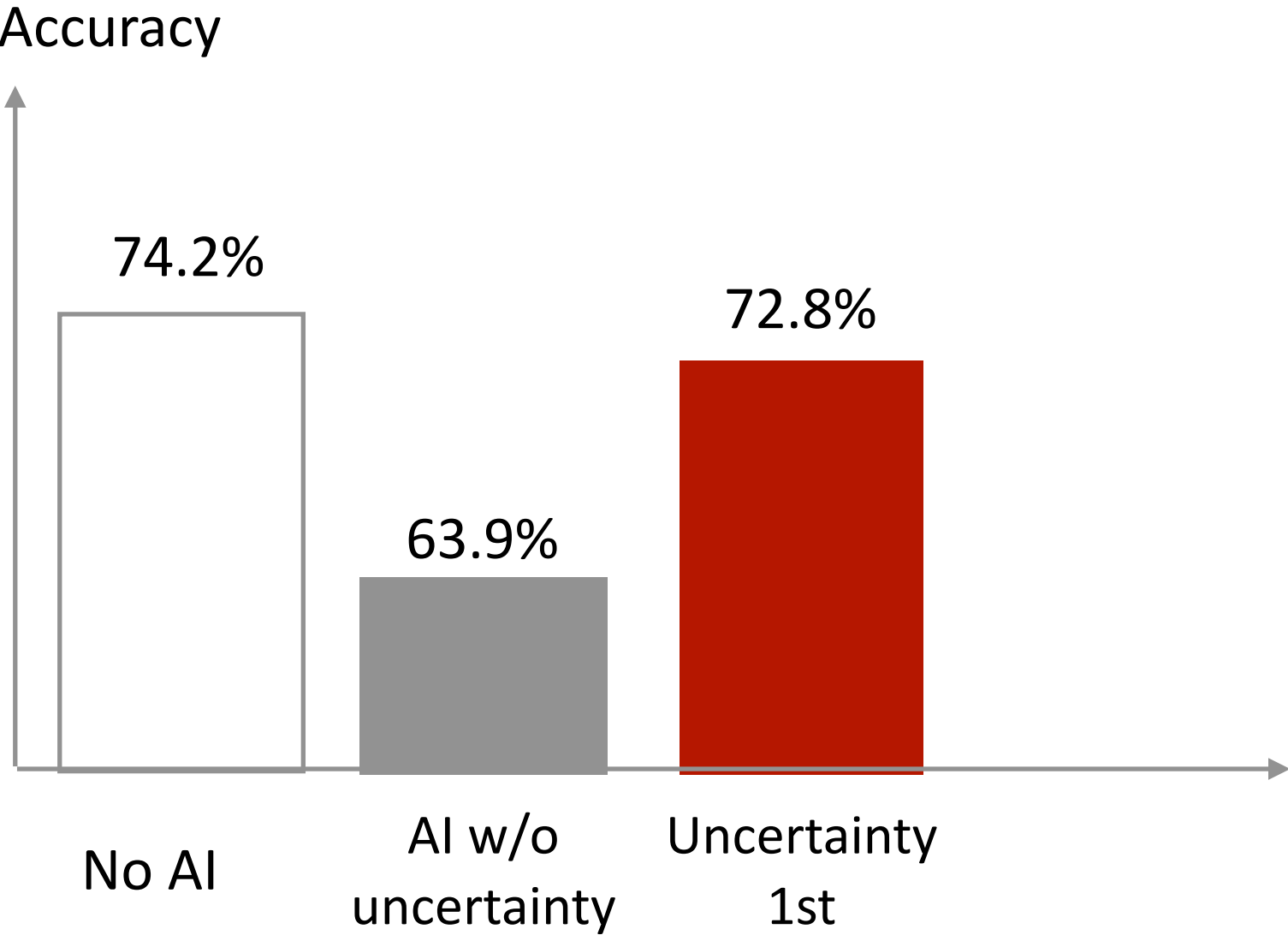
Q. Is Spironolactone an FDA-approved drug for treating acne?

Not uncertain

Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].

Uncertain in the first-person perspective

I'm not sure, but my guess is Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].



Uncertainty expression reduces overreliance, but not eliminate it

Q. Is Spironolactone an FDA-approved drug for treating acne?

Not uncertain

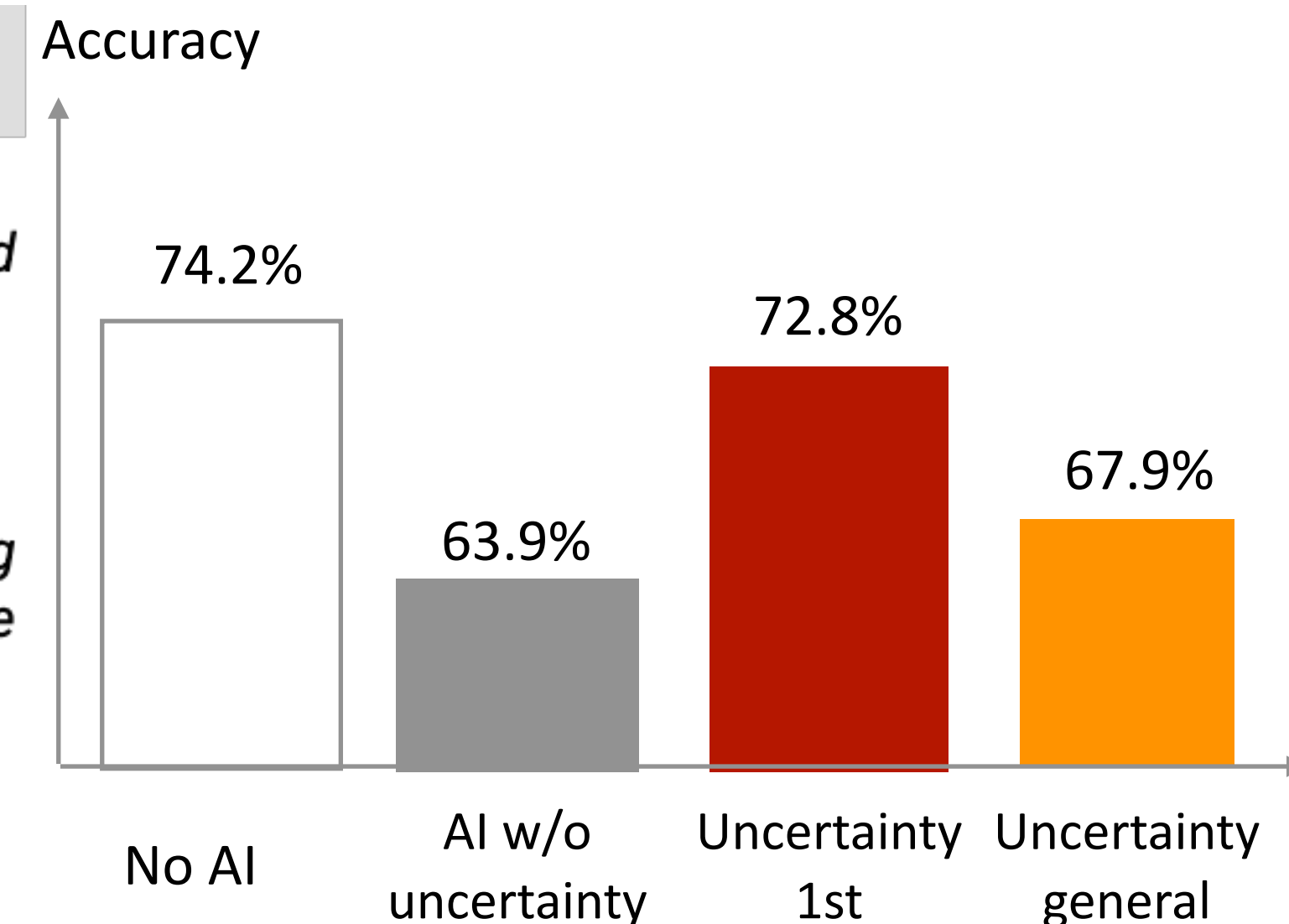
Yes, Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].

Uncertain in the
first-person perspective

I'm not sure, but my guess is Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].

Uncertain in the
general perspective

There is uncertainty, but it seems like Spironolactone is an FDA-approved drug for treating acne. It is primarily used in medicine to treat heart failure and high blood pressure but doctors sometimes prescribe it for acne as well [1, 2, 3, 4].



Uncertainty expression reduces overreliance, but not eliminate it

Linguistic variation matters: general perspective has a weaker effect than first-person expression. But the long-term anthropomorphism risks of the latter call for further investigation

Reducing Conversational Agents' Overconfidence Through Linguistic Calibration

Sabrina J. Mielke^{1,2} Arthur Szlam² Emily Dinan² Y-Lan Boureau²

¹Department of Computer Science, Johns Hopkins University, USA ²Facebook AI Research, USA
sjmielke@jhu.edu {aszlam, edinan, ylan}@fb.com

Abstract

While improving neural dialogue agents' factual accuracy is the object of much research,

control of the alignment between the confidence of an answer and its likelihood of being correct seems like a promising type of remediation: It

Uncertainty in Natural Language Generation: From Theory to Applications

Joris Baan^{1*} Nico Daheim^{2*} Evgenia Ilia^{1*} Dennis Ulmer^{3,4*} Haau-Sing Li²
Raquel Fernández¹ Barbara Plank^{9,3} Rico Sennrich^{7,8} Chrysoula Zerva^{5,6} Wilker Aziz¹

¹University of Amsterdam ²TU Darmstadt & hessian.AI ³IT University of Copenhagen
⁴Pioneer Centre for Artificial Intelligence ⁵Instituto de Telecomunicações
⁶Instituto Superior Técnico & LUMIS (Lisbon ELLIS Unit) ⁷University of Zurich
⁸University of Edinburgh ⁹LMU Munich & Munich Center for Machine Learning

Abstract

Recent advances of powerful Language Models have allowed Natural Language Generation

Touvron et al., 2023a,b). Their widespread use makes it increasingly important to build NLG systems that are trustworthy and representative of the diversity of its users (Bhatt et al., 2021; Joshi et al.,

Uncertainty expression generation needs careful consideration of linguistic characteristics, e.g. in curation of fine-tuning data

Teaching models to express their uncertainty in words

Stephanie Lin
University of Oxford

sylin07@gmail.com

Jacob Hilton
OpenAI

jhilton@openai.com

Owain Evans
University of Oxford

owaine@gmail.com

Abstract

We show that a GPT-3 model can learn to express uncertainty about its own answers in natural language without use of model logits. When given a question, the model generates both an answer and a level of confidence (e.g. “90% confidence” or “high confidence”). These

SEMANTIC UNCERTAINTY: LINGUISTIC INVARIANCES FOR UNCERTAINTY ESTIMATION IN NATURAL LANGUAGE GENERATION

Lorenz Kuhn, Yarin Gal, Sebastian Farquhar
OATML Group, Department of Computer Science, University of Oxford
lorenz.kuhn@cs.ox.ac.uk

ABSTRACT

We introduce a method to measure uncertainty in large language models. For tasks like question answering, it is essential to know when we can trust the natural language outputs of foundation models. We show that measuring uncertainty in natural language is challenging because of ‘semantic equivalence’—different sentences can mean the same thing. To overcome these challenges we introduce

Failures contributing to socio-technical gap of AI

Failure in accounting for human interaction

Failure in anticipating new and diverse use cases

Failure in formalizing “socio” requirements

How can HCI research help address these failures?

Empirical evaluation

Formative research

Theory and framework building

RAI's “Failure of imagination” Problem

Overcoming *Failures of Imagination* in AI Infused System Development and Deployment

Margarita Boyarskaya
New York University

Alexandra Olteanu
Microsoft Research

Kate Crawford
Microsoft Research

Abstract

NeurIPS 2020 requested that research paper submissions include impact statements on “potential nefarious uses and the consequences of failure.” However, as researchers, practitioners and system designers, a key challenge to anticipating risks is overcoming what Clarke (1962) called ‘failures of imagination.’ The growing research on bias, fairness, and transparency in computational systems aims to illuminate and mitigate harms, and could thus help inform reflections on possible negative impacts of particular pieces of technical work. The prevalent notion of *computational harms*—narrowly construed as either allocational or representational harms—does not fully capture the open, context dependent, and unobservable nature of harms across the wide range of AI infused systems. The current literature focuses on a small range of examples of harms to motivate algorithmic fixes, overlooking the wider scope of probable harms and the way these harms might affect different *stakeholders*. The *system affordances* may also exacerbate harms in unpredictable ways, as they determine stakeholders’ control (including of non-users) over how they use and interact with a *system output*. To effectively assist in anticipating harmful uses, we argue that frameworks of harms must be context-aware and consider a wider range of potential *stakeholders*, *system affordances*, as well as viable proxies for *assessing harms* in the widest sense.

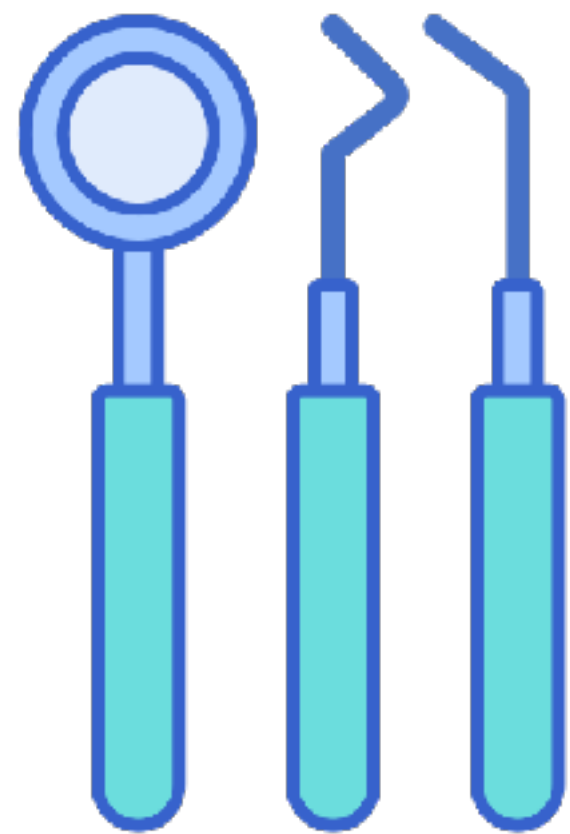
Researchers’ personal biases on what use cases matter

Application developers’s adoption without guidance

End-users’ appropriation of technology

Changes of the ecosystem

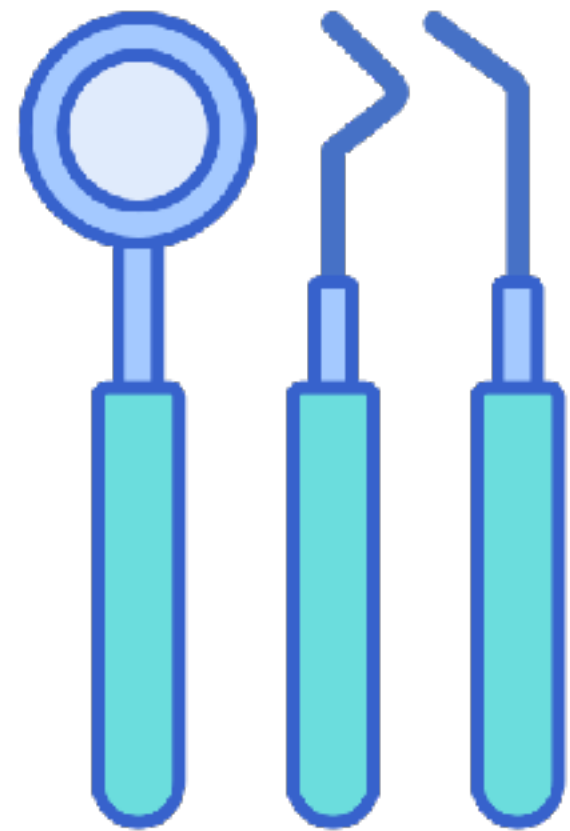
Probe-Based Needs-Finding Study



Technology probe: an instrument that is deployed to find out about the unknown - to hopefully return with useful or interesting data

Transparency probe: an instrument that represents capability of current transparency approach(es), studied in a realistic use context, to find out about the gaps in understanding

What Can a Transparency Probe Do?



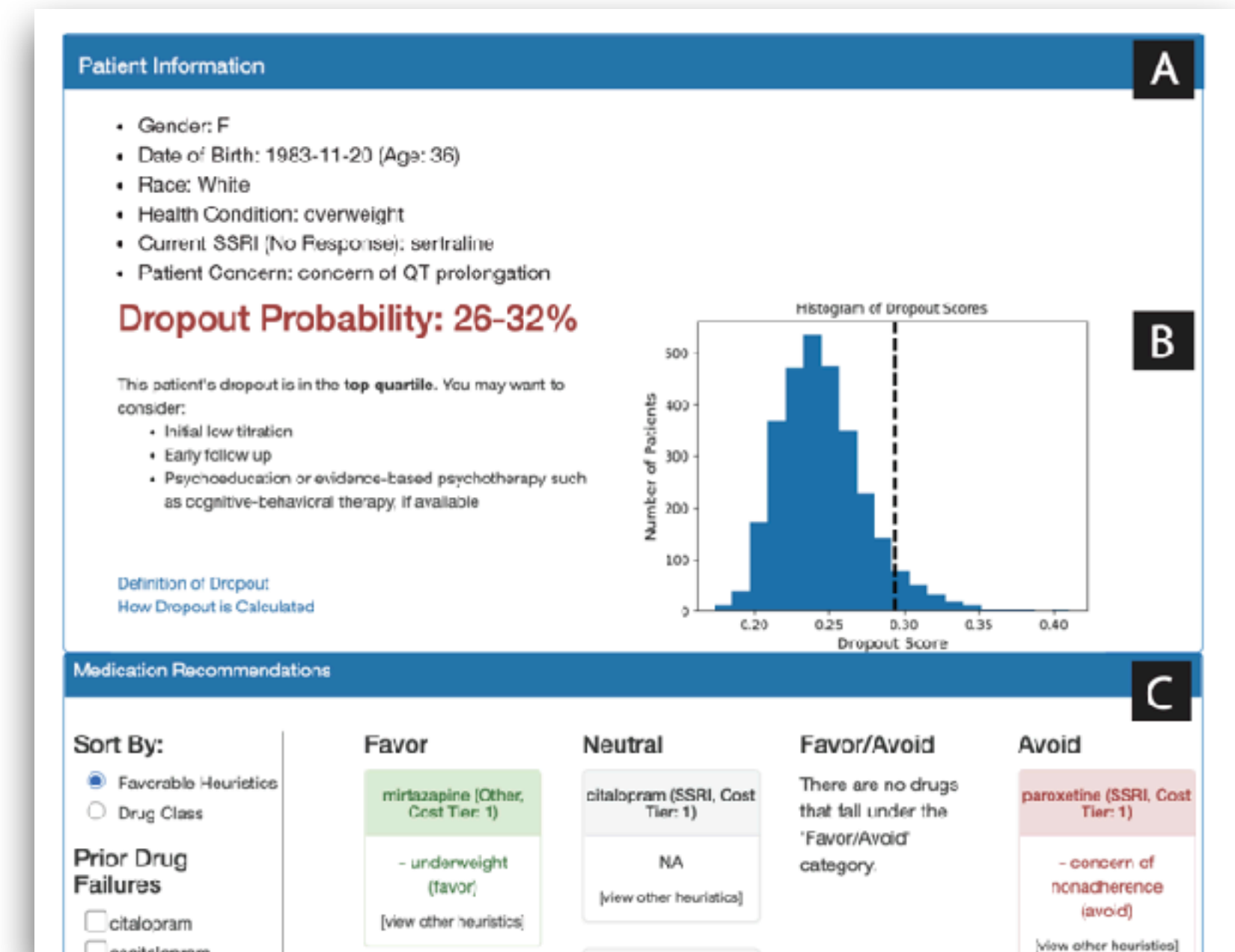
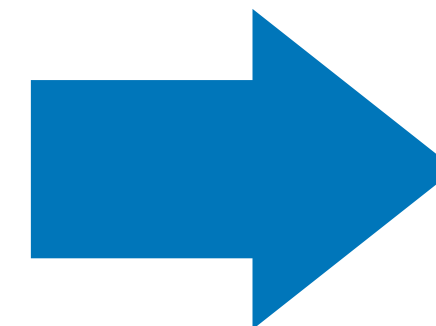
- **Technical goal:** concretely assess what the current transparency approach can or cannot do
- **Social science goal:** understand users' transparency goals and needs in a realistic setting
- **Design goal:** inspire new transparency approaches with user input

What Can Be a Transparency Probe?



A low-fi prototype with a patient scenario to elicit physician needs' for explanations from medical decision support tool

What Can Be a Transparency Probe?



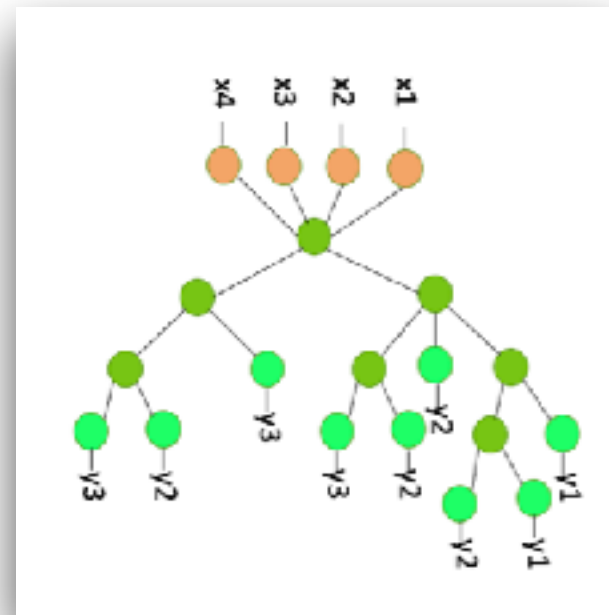
A low-fi prototype with a patient scenario to elicit physician needs' for explanations from medical decision support tool

Understand system constraints

Do not require determination of trust at every decision point

- Refocus from model features to model validation process
- Present an overview of all model validation steps in a single screen that is accessible from the main interface, but not combined with patient details

Transparency Probes: Questions Answerable by Current XAI Techniques



Global explanation: *How* does the model make decisions?



Feature-importance: *Why* is this instance predicted to be negative?

•If {debt percentage under 30%},
you will no longer be
predicted of high risk

Counterfactual explanation: *How to* get a *different* prediction?

Understanding the model globally: How does the system make predictions (overall logic)?

- What algorithm is used?
- What *rules* does the system use to make predictions?
- *What features* does the model consider or not consider?
- How does the model *weigh/reason with these features*?

Understanding prediction for a particular case: Why this? Why not that?

- Why is this case given this prediction?
- What *feature(s)* of this case lead to the prediction?
- *What kind of cases* are predicted this?
- Why are [cases A and B] given *the same* prediction?
- Why are [cases A and B] given *different* predictions?

Inspecting what if changing a case/counterfactual questions: what if, how to be that, how to still be this

- What would the system predict if the case changes to...?
- How should this case change to get a different prediction?
- What are the scope of changes permitted for this case to still get the same prediction?
- What kind of cases get a different/same prediction?

Other category (add your own question)

Opportunities for New XAI Approaches

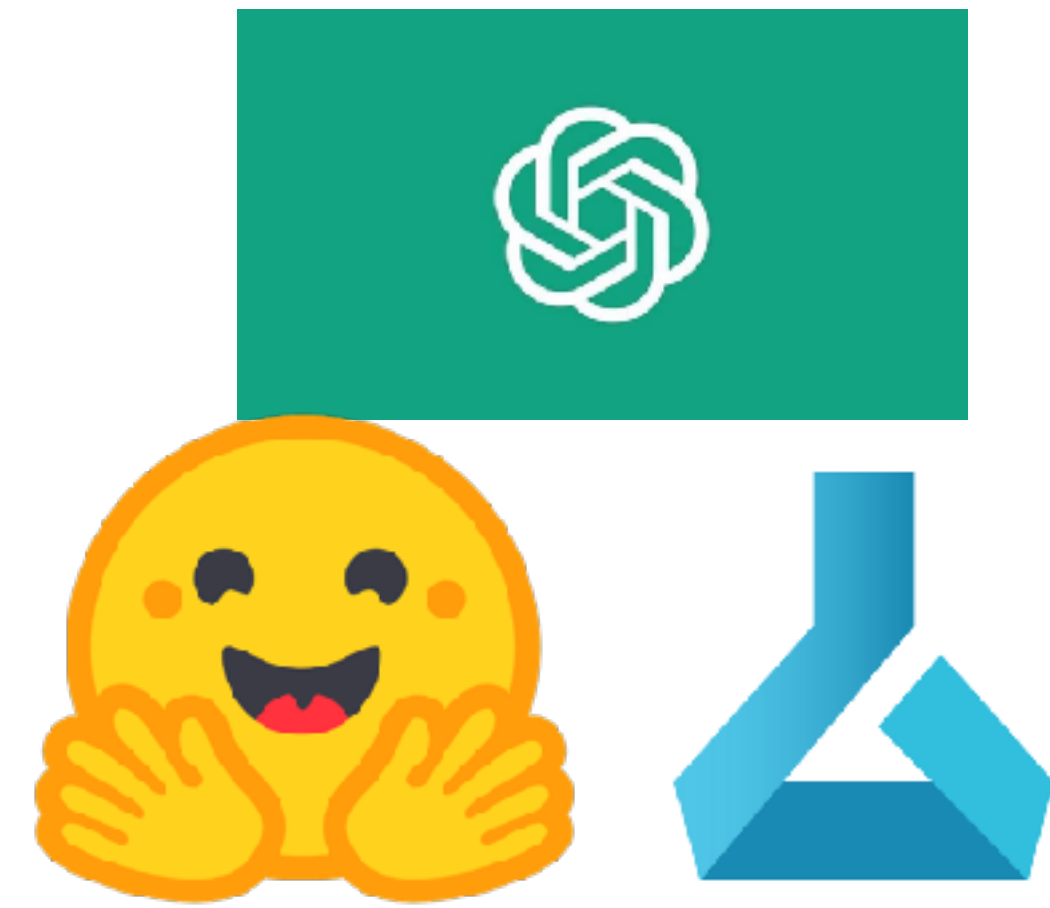
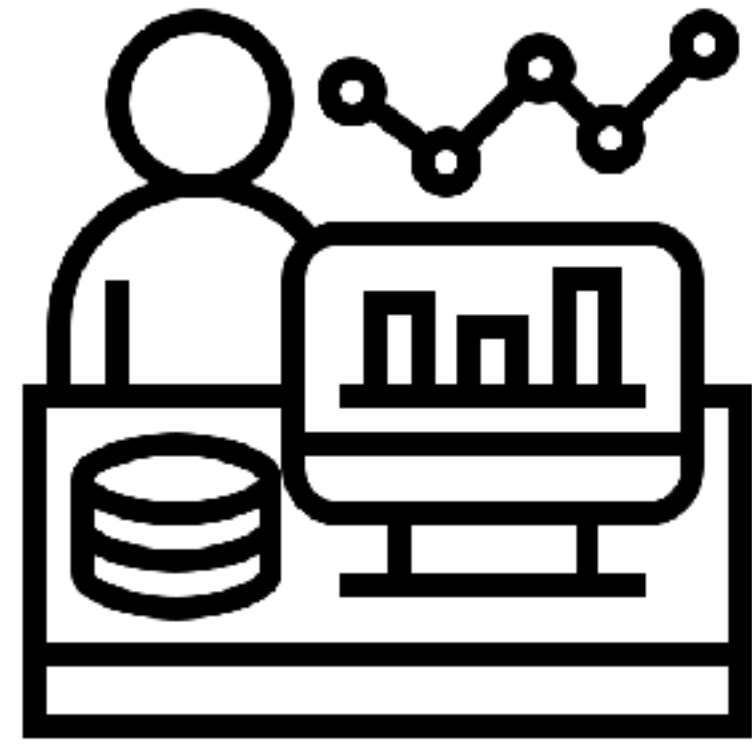
What user questions cannot be answered?

- Explain data limitations and generalizability
- Explain output of multiple models
- Explain system changes
- Multi-level global explanations
- Interactive counterfactual explanations
- Social explanations
- ...

Transparency Probe: Model Card

Model Card
• Model Details. Basic information about the model. – Person or organization developing model – Model date – Model version – Model type – Information about training algorithms, parameters, fairness constraints or other applied approaches, and features – Paper or other resource for more information – Citation details – License – Where to send questions or comments about the model • Intended Use. Use cases that were envisioned during development. – Primary intended uses – Primary intended users – Out-of-scope use cases • Factors. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3. – Relevant factors – Evaluation factors • Metrics. Metrics should be chosen to reflect potential real-world impacts of the model. – Model performance measures – Decision thresholds – Variation approaches • Evaluation Data. Details on the dataset(s) used for the quantitative analyses in the card. – Datasets – Motivation – Preprocessing • Training Data. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets. • Quantitative Analyses – Unitary results – Intersectional results • Ethical Considerations • Caveats and Recommendations

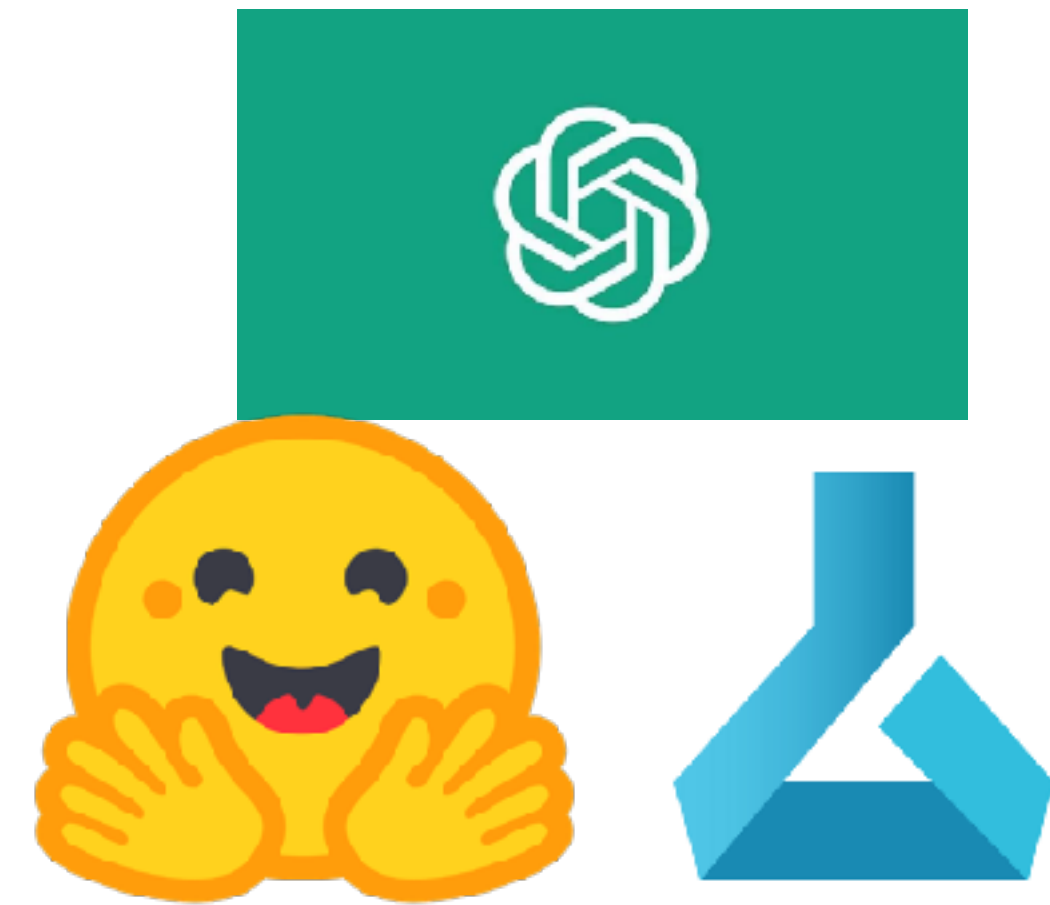
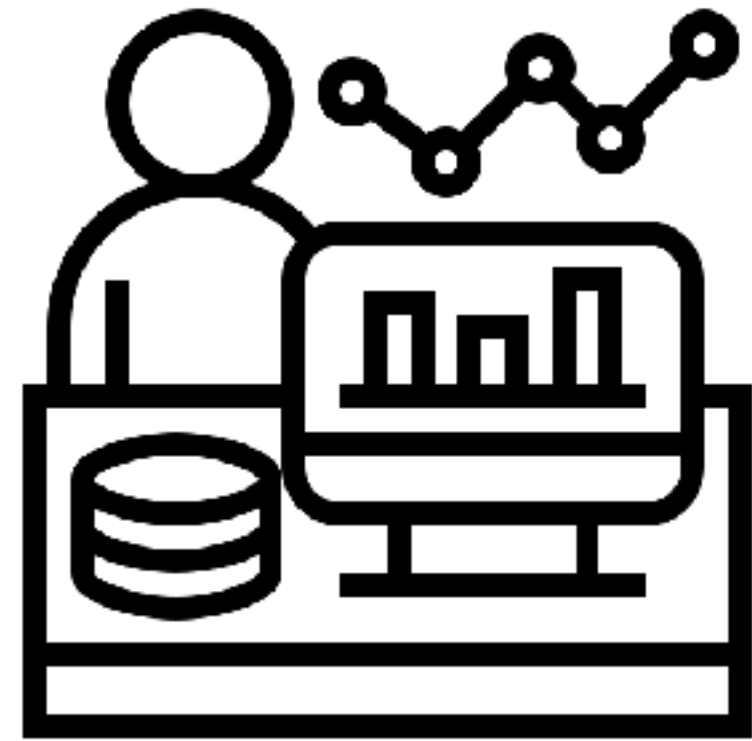
Changes of Ecosystem and Use Cases



2019

2022

Changes of Ecosystem and Use Cases



Non-technical practitioners (e.g. designers) frequently need to **understand pre-trained models to perform design ideation**

2019

2022

Method: Probe-Based Needs-Finding



Design task to use a pre-trained model

Model description: the basics of extractive summarization

The extractive summarization model uses natural language processing techniques to locate key sentences in an unstructured text document. These sentences collectively convey the main idea of the document.

When a document is given as the input, the model returns a list of extracted sentences, together with a rank score and its position in the original document for each extracted sentence. A rank score is an indicator of how relevant or important the model considers the sentence is to the main idea of the document (between 0 and 1, higher as more relevant).

By default, the model returns three highest scored sentences, and you can specify the number of sentences returned.

Examples of intended use

You might want to use the extractive summarization model to:

- Distill critical information from lengthy documents.
- Highlight key sentences in documents.
- Quickly skim documents in a library.
- Generate news feed content.

You can use extractive summarization in multiple scenarios across a variety of industries. For example:

- Extract key information from public news articles to produce insights.
- Classify documents by their key contents.
- Distill important information from long documents to empower solutions such as search, question and answering, and decision support.

Do not use

Don't use extractive summarization for automatic actions without human intervention for high-impact scenarios. A person should always review source data when another person's economic situation, health, or safety is affected.

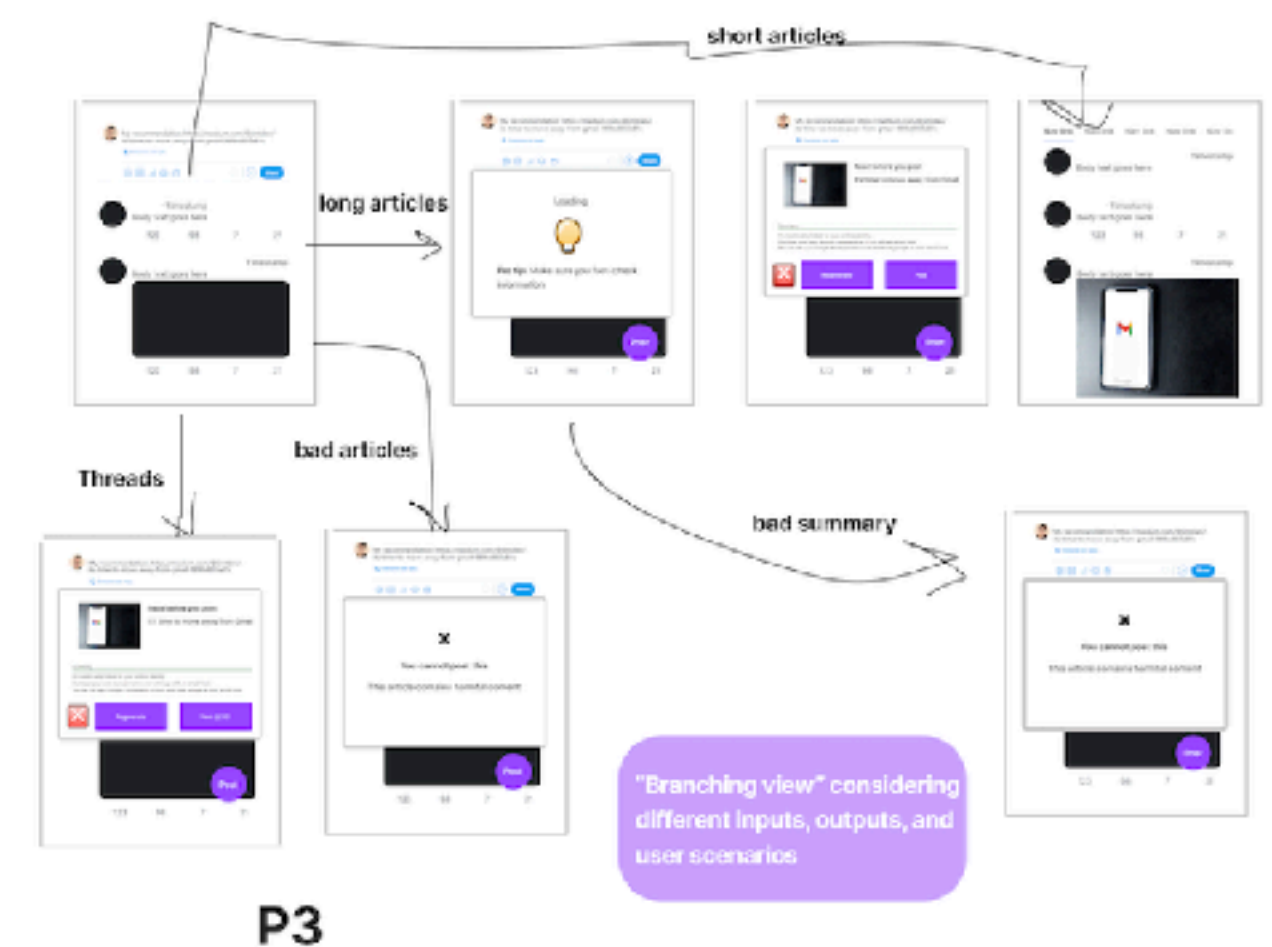
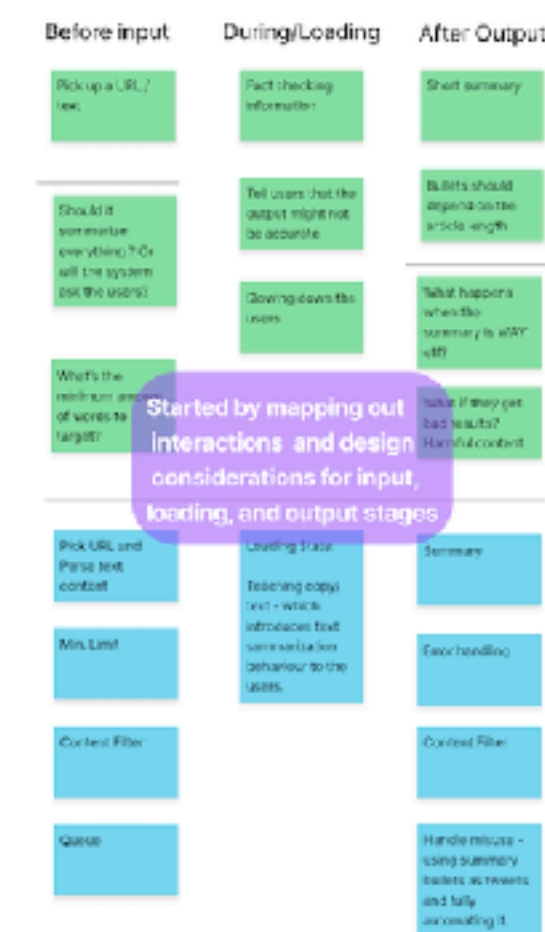
Limitations with impacting factors

Based on your scenario and input data, you could experience different levels of performance:

- Because the model is trained on document-based texts, such as news articles, scientific reports, and legal documents, when used with texts in certain genres such as conversations and transcriptions, it might produce output with lower accuracy.
- When used with texts that may contain errors or are less similar to well-formed sentences, such as texts extracted from lists, tables, charts, the model might produce output with lower accuracy.

Model Documentation

- **Model description**
 - Descriptions of input and output
 - Who, when and how it was developed
 - Model type and algorithm
 - Training data
- **Intended use**
 - Examples
 - Do not use
- **Limitations with impacting factors**
 - What factors affect model performance
- **Performance evaluation**
 - Evaluation metrics
 - Evaluation methods and data
- + Design space guidance?
- + Potential harms and ethical considerations
- + ...



Read a model card

Perform ideation

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms consideration, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and information about user contexts, such as providing or helping designers to define user workflows or scenarios. - Scaffold defining UX risks by discovering a broad range of model limitations and ethical consideration guidance. - Support risk-benefit assessment of candidate ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors in limitations and harms consideration, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms consideration	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or user needs, such as restructuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms consideration, limitations, design space guidance	customizability and improvability, algorithm, , training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

What did we learn?

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms consideration, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and information about user contexts, such as providing or helping designers to define user workflows or scenarios. - Scaffold defining UX risks by discovering a broad range of model limitations and ethical consideration guidance. - Support risk-benefit assessment of candidate ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors on limitations and harms consideration, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms consideration	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or user needs, such as restructuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms consideration, limitations, design space guidance	customizability and improvability, algorithm, , training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

What did we learn?

- **Technical goal:** what current model card can or cannot do

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms consideration, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and information about user contexts, such as providing or helping designers to define user workflows or scenarios. - Scaffold defining UX risks by discovering a broad range of model limitations and ethical consideration guidance. - Support risk-benefit assessment of candidate ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors in limitations and harms consideration, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms consideration	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or user needs, such as restructuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms consideration, limitations, design space guidance	customizability and improvability, algorithm, , training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

What did we learn?

- **Technical goal:** what current model card can or cannot do
- **Social science goal:** what are designers goals with understanding pre-trained models?

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms consideration, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and information about user contexts, such as providing or helping designers to define user workflows or scenarios. - Scaffold defining UX risks by discovering a broad range of model limitations and ethical consideration guidance. - Support risk-benefit assessment of candidate ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors in limitations and harms consideration, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms consideration	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or user needs, such as restructuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms consideration, limitations, design space guidance	customizability and improvability, algorithm, , training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

What did we learn?

- **Technical goal:** what current model card can or cannot do
- **Social science goal:** what are designers goals with understanding pre-trained models?
- **Design goal:** inspire new approaches to support designedly understanding

Transparency goal	Provided info used	Requested information	Design implications to support the goal
G1: Divergent-convergent design thinking	intended uses, model description, input-output examples, harms consideration, unintended uses, limitations	output analysis, explanations	- Inspire divergent design thinking with example use cases and information about user contexts, such as providing or helping designers to define user workflows or scenarios. - Scaffold defining UX risks by discovering a broad range of model limitations and ethical consideration guidance. - Support risk-benefit assessment of candidate ideas; develop risk-oriented evaluation metrics and practices.
G2: Conditional design	impacting factors in limitations and harms consideration, examples of different categories	training data, explanations, disaggregated evaluation with performance and other output characteristics, confidence/uncertainty	- Facilitate discovering and testing hypotheses about impacting factors and assessing their impact. - Support decisions about whether to create conditional designs by exploring the design space (e.g., creating intermediate prototypes) and assessing user values. - Develop design patterns and implementation guidance for conditional designs.
G3: Transparency for users	model description, limitations, harms consideration	performance, confidence/uncertainty, explanations	- Allow seeking model information by user questions or user needs, such as restructuring the documentation and providing other information channels. - Support translating information in documentation to transparency designs for users.
G4: Team negotiation and collaboration	harms consideration, limitations, design space guidance	customizability and improvability, algorithm, , training data and other development information	- Empower designers by prioritizing their suitability assessment of the model for the users, with both informational and organizational support. - Equip designers to collaborate with engineers with technical literacy, actionable suggestions for model improvement, common references, and boundary objects.

Transparency goal:
convergent design thinking
by eliminating “risky” ideas

Design implications:
scaffold discovering model
failures in UX context,
defining metrics and
quantifying the failure

Designerly Understanding of Model Limitations with fAllureNotes

a

User Personas and Scenarios

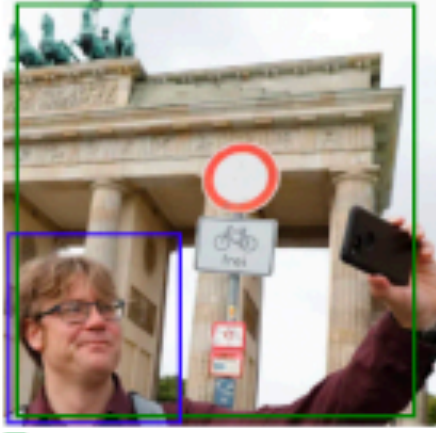
Tom, the city tourist
Tom enjoys learning a new language while traveling to vibrant cities.

Tom is on a guided walking tour in Berlin.

Tom is eating in a local restaurant.

b

User Expectation



Brandenburger Gate

Person

Save Start over

Annotating the desired model behavior (user expectation)

c

Model Prediction



Person 99%

Cell Phone 99%

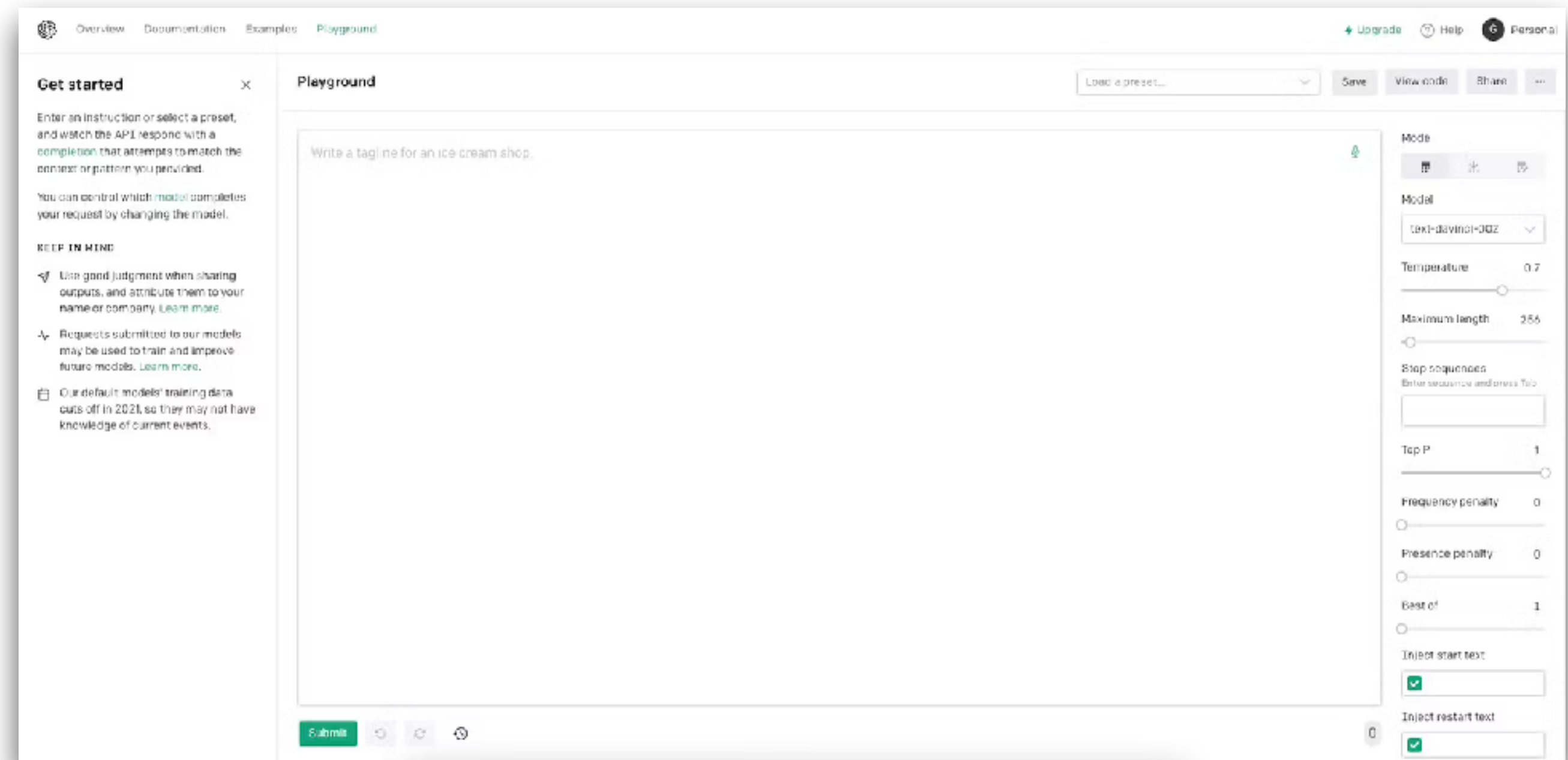
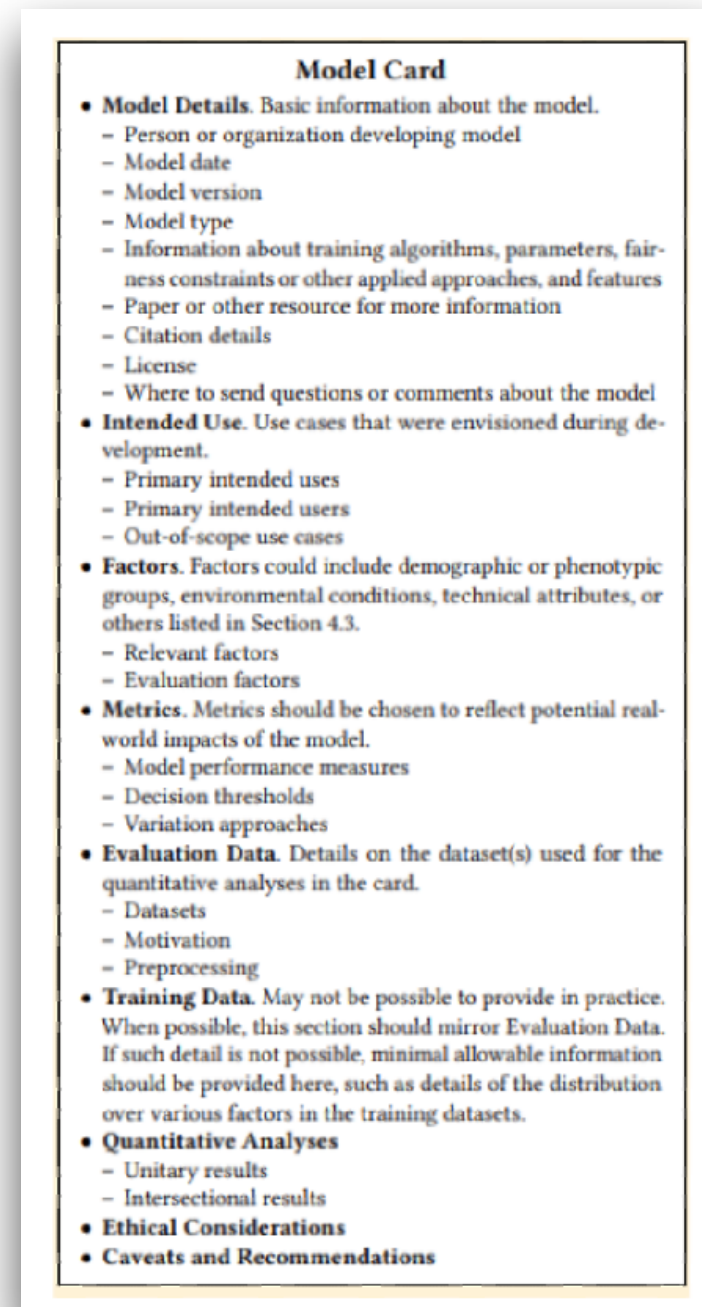
Failure Engine

#	User Expectation	Model Prediction	Failure Mode	Failure Type	Warnings	Severity (1 low - 7 high)
1	Person 99%	Person	CD	ID	CQB	
2	Brandenburger Gate		MD	OOD		4
3		Cell Phone 99%	UD			1

Save

Exploring mismatch between expected and actual model behavior

FailureNotes: Scaffolding designers's discovery and assessment of UX specific model failures



How to provide transparency in the playground UI of pre-trained models for diverse use cases?

2019

2025

Failures contributing to socio-technical gap of AI

Failure in accounting for human interaction

Failure in anticipating new and diverse use cases

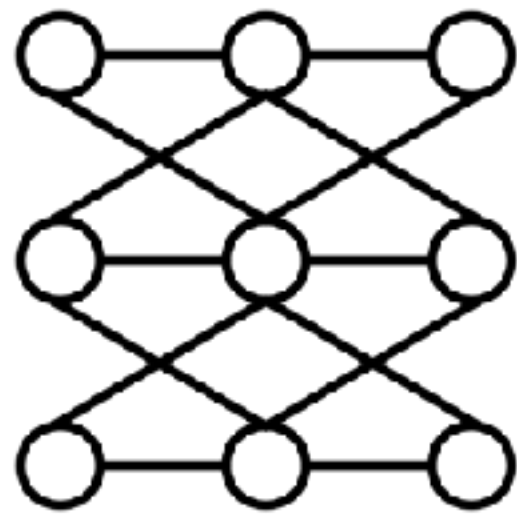
Failure in formalizing “socio” requirements

How can HCI research help address these failures?

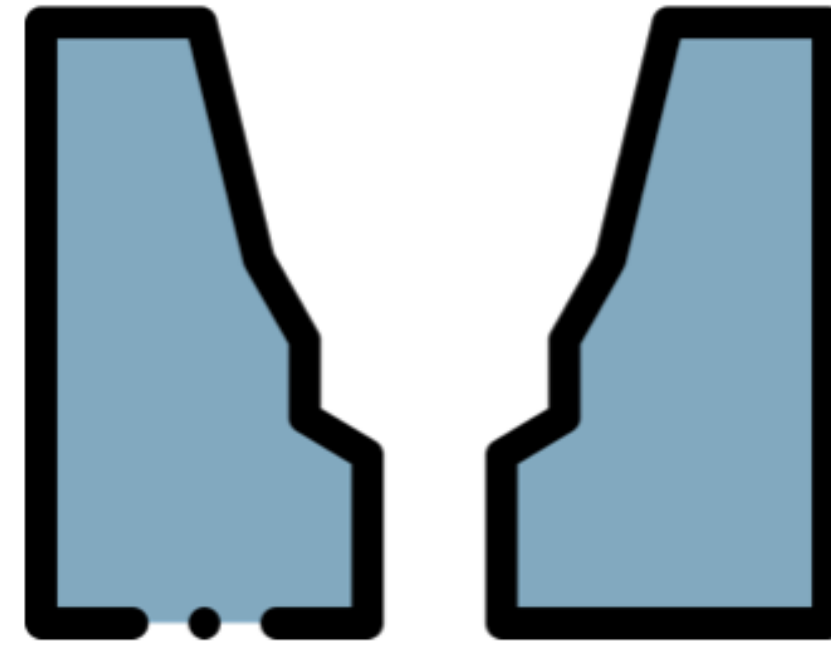
Empirical evaluation

Formative research

Theory and framework building



Precise, formal, abstract...



Contextual, nuanced, fluid...



How to formalize contextual and nuanced human requirements rigorously, grounded in empirical data, in what AI optimizes for?

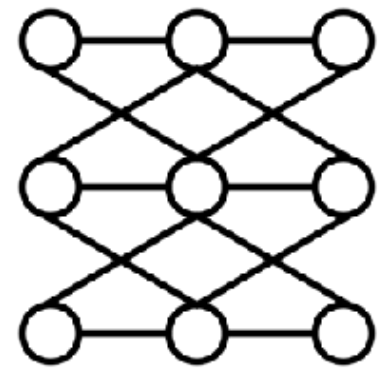
A Bridging Point: AI Evaluation



Evaluation sets the optimization goals and shapes technical development

Evaluation *is* the formalization of what is a “good” AI technology

XAI Evaluation Gap



**XAI evaluation
criteria**

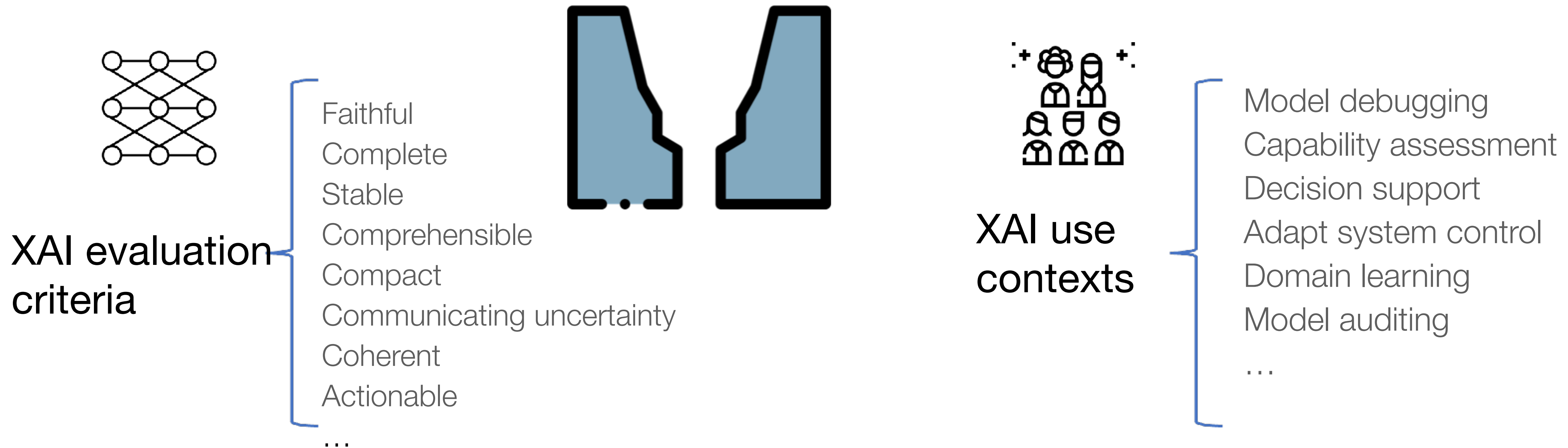
Faithful
Complete
Stable
Comprehensible
Compact
Communicating uncertainty
Coherent
Actionable

...

XAI Evaluation Gap

Do these criteria matter to people?

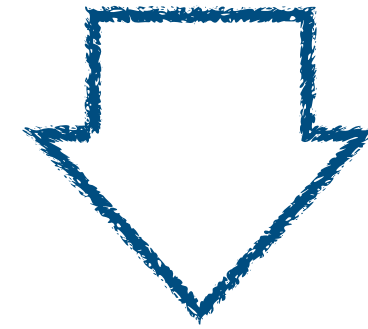
Do these criteria matter to people in different use contexts?



XAI Evaluation Gap

Do these criteria matter to people?

Do these criteria matter to people in different use contexts?



What are the different use contexts of XAI?
...formalized at a meaningful abstraction

Framework of XAI Design Space by Transparency Goals

<i>Goals & Objectives</i>	<i>Phases of the ML Lifecycle where Interpretability Objectives Occur</i>			
	Development	Deployment	Immediate Usage	Downstream Impact
G1: Understanding				
G2: Trust				
O1: Debug & improve				
O2: Compliance w/ regulations				
O3: Act based on output				
O4: Justify actions				
O5: Understand data usage				
O6: Learn about a domain				
O7: Contest decision				

Scenario-Based Survey to Elicit Context-Specific Importance of Evaluation Criteria

Scenario: InvestAssistant is a new AI system that assists personal finance by recommending investment plans (e.g. stocks to sell and buy). The AI system monitors the financial market and considers a user's financial profile and personal preferences such as risk tolerance in order to make recommendations. It learns from historical investment records.

The AI system can generate explanations, e.g. how it judges the risk and return of investment plans, and why it recommends a particular investment plan.

Task 1 of 5

You are an investor and were just introduced to the AI system. While testing the system, you request explanations understand its risks and limitations, so that you could decide whether and how to use the system; e.g. what kind of investments the AI might underestimate the risk.

	Not important	Slightly important	Moderately Important	Important	Very important
Translucence: the explanation is transparent about its limitations, for examples the conditions for it to hold	☐	☐	☐	☐	☐
Coherence: the explanation is consistent with what I already know about the domain	☐	☐	☐	☐	☐

Formalizing “Socio” Requirements: Use-Informed Contextualized Evaluation

Model Improve.		Capability Assess.		Decision Support		Adapt Control		Domain		Model Auditing		Overall	
Expert Survey													
1. Faithfulness	4.61	Translucence	4.40	Interactivity	4.29	(Un)Certainty	4.31	Translucence	4.23	Faithfulness	4.54	Faithfulness	4.30
2. Stability	4.27	(Un)Certainty	4.29	(Un)Certainty	4.26	Interactivity	4.31	Faithfulness	4.20	Translucence	4.43	Translucence	4.25
3. (Un)certainty	4.24	Comprehend.	4.20	Faithfulness	4.14	Translucence	4.29	Stability	4.17	Completeness	4.37	(Un)Certainty	4.15
4. Translucence	4.06	Faithfulness	4.09	Translucence	4.11	Faithfulness	4.26	Comprehen.	4.14	(Un)Certainty	4.00	Stability	4.01
5. Comoprehen.	3.94	Stability	4.00	Comprehend.	3.97	Stability	4.00	Completeness	4.00	Interactivity	3.89	Interactivity	3.92
6. Completeness	3.67	Compactness	4.00	Actionability	3.91	Actionability	3.86	Interactivity	3.86	Stability	3.83	Comprehend.	3.89
7. Actionability	3.48	Interactivity	3.80	Stability	3.83	Completeness	3.80	(Un)Certainty	3.83	Comprehend.	3.49	Completeness	3.70
8. Compactness	3.42	Actionability	3.69	Personalization	3.43	Comprehen.	3.63	Compactness	3.60	Actionability	3.06	Actionability	3.55
9. Interactivity	3.33	Personalization	3.63	Compactness	3.31	Coherence	3.46	Personalization	3.57	Compactness	3.03	Compactness	3.45
10. Coherence	3.18	Coherence	3.54	Novelty	3.20	Compactness	3.31	Actionability	3.29	Coherence	2.97	Coherence	3.23
11. Novelty	2.88	Completeness	3.23	Coherence	3.17	Personalization	2.94	Novelty	3.17	Personalization	2.57	Personalization	3.14
12. Personalizatio	2.67	Novelty	2.51	Completeness	3.11	Novelty	2.91	Coherence	3.03	Novelty	2.37	Novelty	2.82

Contextualized evaluation: evaluate by prototypical downstream use contexts with context-specific evaluation criteria

- *Responsible research* by articulating intended and unintended use
- Support practitioners and end-users’ assessment for their context



The socio-technical gap of AI is widening in the age of “large” models

The evaluation gap is also widening, often failing predict the utility of “large” models in diverse real-world contexts

Holistic Evaluation of Language Models

Percy Liang[†], Rishi Bommasani[†], Tony Lee[†], Dimitris Tsipras[‡], Dilara Soylu[‡], Michihiro Yasunaga[‡], Yian Zhang[‡], Deepak Narayanan[‡], Yuhuai Wu[‡], Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, Yuta Koreeda

pliang@cs.stanford.edu, rlbprishi@stanford.edu, tonyhlee@stanford.edu

Center for Research on Foundation Models (CRFM)

Institute for Human-Centered Artificial Intelligence (HAI)

Stanford University

Reviewed on OpenReview: <https://openreview.net/forum?id=i04LZibEqW>

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Alphabetic author list:*

AaroHi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shueb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubakaran, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinon, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin

GenAI Evaluation Crisis

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

ABSTRACT

We propose a new test to measure a text model’s multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more. To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability. We find that while most recent models have near random-chance accuracy, the very largest GPT-3 model improves over random chance by almost 20 percentage points on average. However, on every one of the 57 tasks, the best models still need substantial improvements before they can reach expert-level accuracy. Models also have lopsided performance

Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference

Wei-Lin Chiang^{*1} Lianmin Zheng^{*1} Ying Sheng² Anastasios N. Angelopoulos¹ Tianle Li¹ Dacheng Li¹
Banghua Zhu¹ Hao Zhang³ Michael I. Jordan¹ Joseph E. Gonzalez¹ Ion Stoica¹

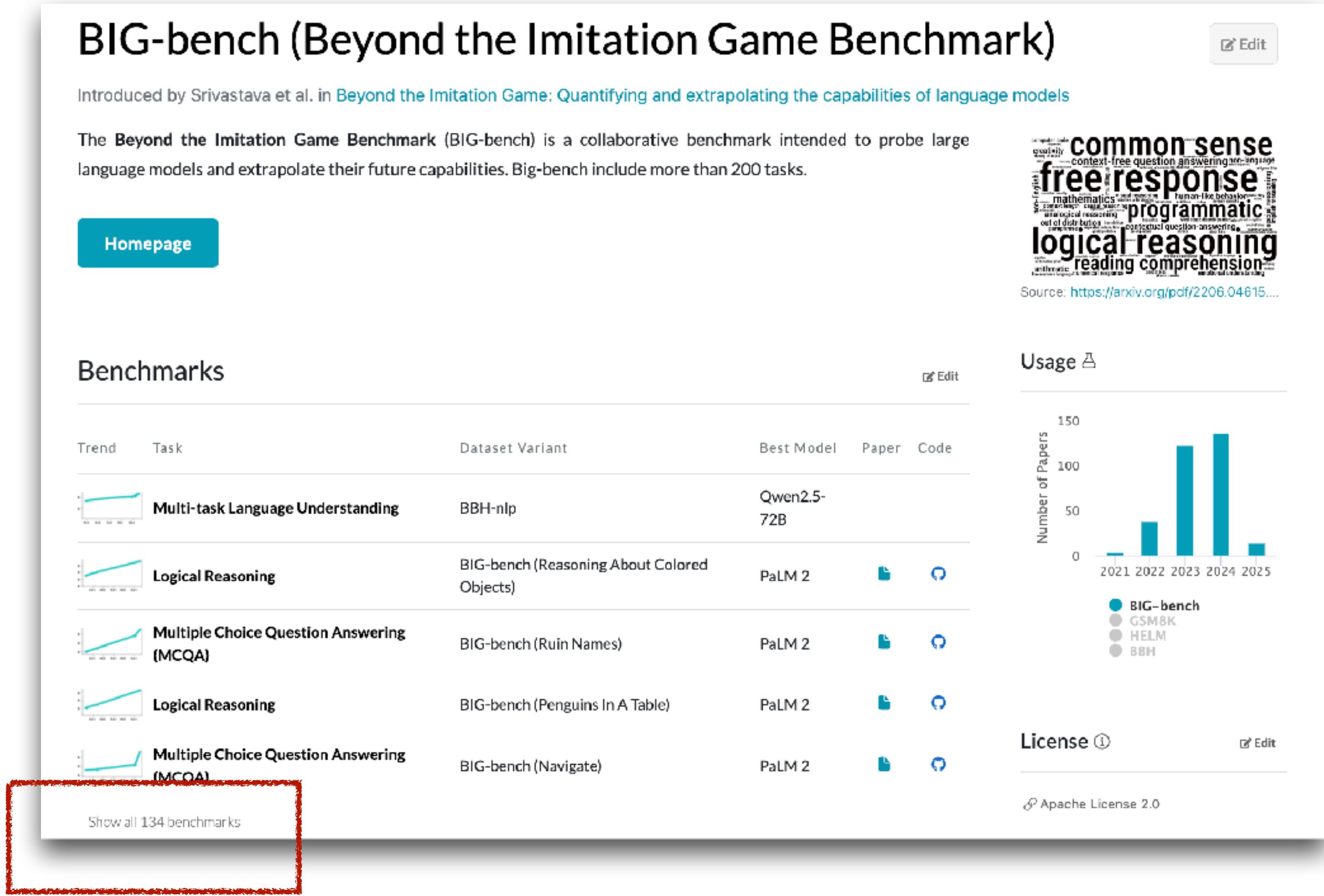
Abstract

Large Language Models (LLMs) have unlocked new capabilities and applications; however, evaluating the alignment with human preferences still poses significant challenges. To address this issue, we introduce Chatbot Arena, an open platform for evaluating LLMs based on human preferences. Our methodology employs a pairwise comparison approach and leverages input from a diverse user base through crowdsourcing. The platform has been operational for several months, amassing over 240K votes. This paper describes

Evaluation Metric	Question Source	
	Static	Live
	Ground Truth	MMLU, HellaSwag, GSM-8K
Human Preference	MT-Bench, AlpacaEval	Codexforces Weekly Contests
		Chatbot Arena

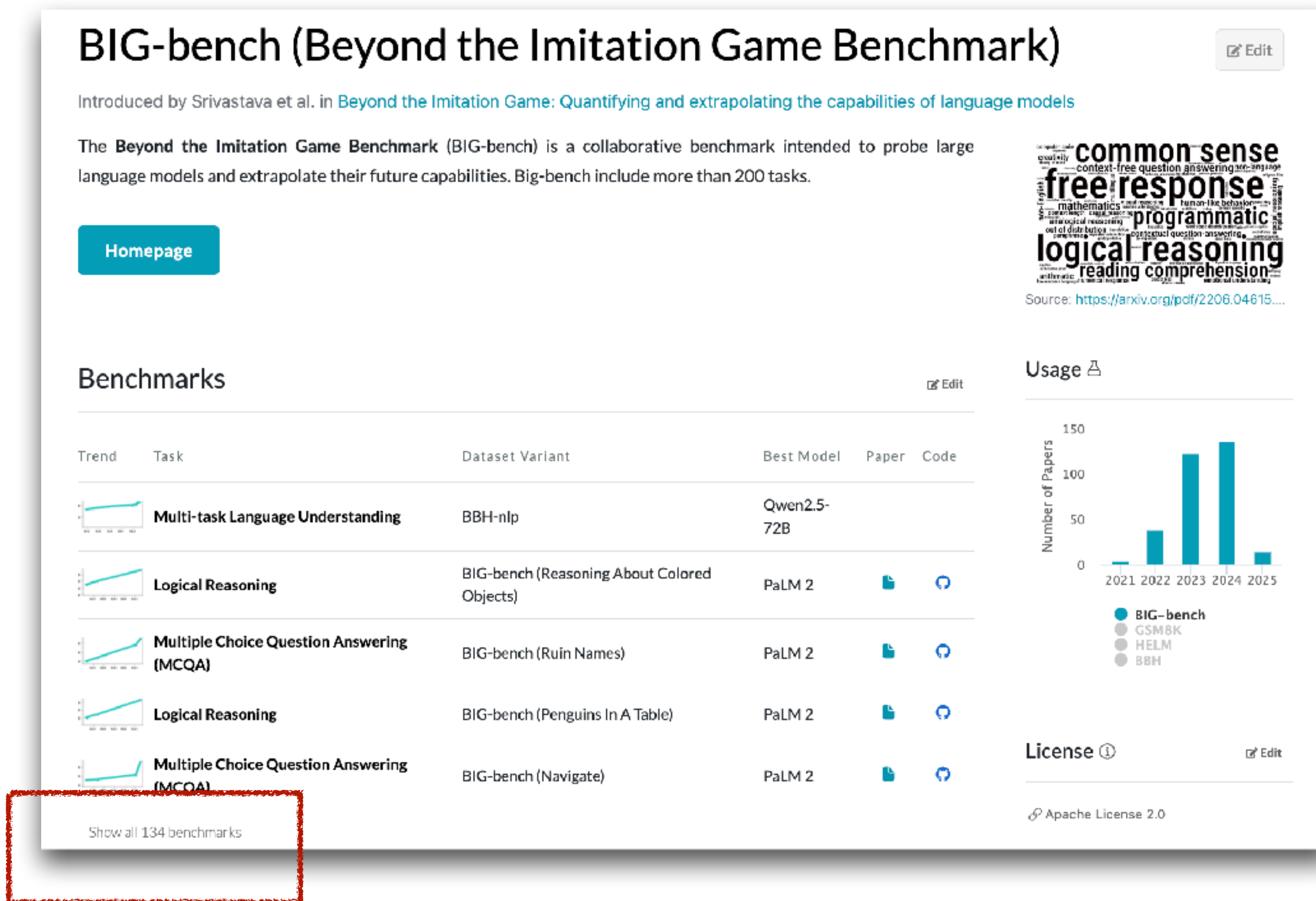
Figure 1. Classification of LLM benchmarks: We categorize along two dimensions: whether the questions are from a static dataset or a live, fresh source, and whether the evaluation metric relies on ground truth or (approximated) human preferences. MMLU (Hendrycks et al., 2020), HellaSwag (Zellers et al., 2019), GSM-8K (Cobbe et al., 2021), MT-Bench (Zheng et al., 2023b), and AlpacaEval (Li et al., 2023) are common examples of static benchmarks. Chatbot Arena is the platform introduced in this paper.

Current Trend: Meta-Benchmarks



Srivastava et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models

Current Trend: Meta-Benchmarks



- Lack validity: benchmarks do not map to real-world concepts that matter to the use and impacts of GenAI
- Lack meaningful organization: difficult for stakeholders to know what to look for

Current Trend: Taxonomizing the Evaluation Space

Use cases (scenario)

Criteria (disederata)

	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✓	✓	✓	✓	✓	✓	✓
IMDB	✓	✓	✓	✓	✓	✓	✓
Natural Questions	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓
XSUM	✓				✓	✓	✓

Define the Evaluation Space by Formalizing “Socio” Requirements

What are the context-specific human requirements?

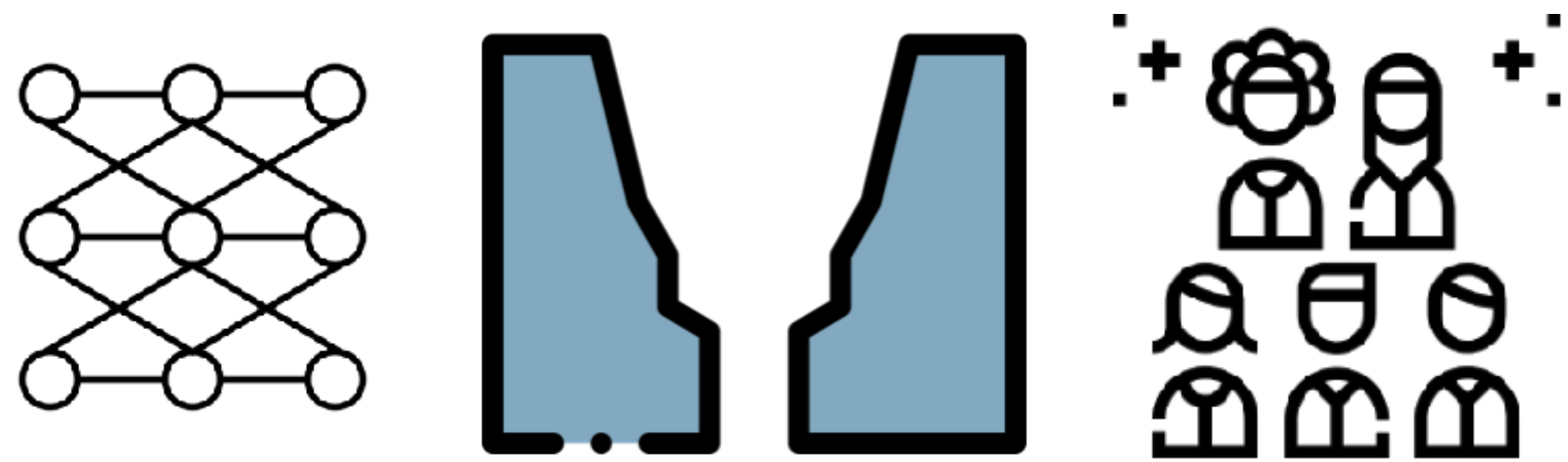
Evaluation Criteria

How to design benchmark tasks and metrics to realistically reflect the use contexts and human impacts?

What are the use contexts?

Use Contexts

Bridging the Socio-Technical Gap of AI: Final Thoughts



Critical investigation

Rigorous formalization

Iterations through the
ecosystem

Building communities at the
connection points



H CX AI

HUMAN-CENTERED EXPLAINABLE AI

logo by [Upol Ehsan](#)

HEAL @ CHI 2025

Human-centered Evaluation and Auditing of Language Models

Yokohama, Japan | April 26, 2025

Thank **YOU!**

And thanks to my amazing collaborators:

Alexandra Olteanu, Alison Smith-Renner, Ameneh Shamekhi, Amit Dhurandhar, Anna Kawakami, Bart P.Knijnenburg, Bhavya Ghai, Biplav Srivastava, Bruce Schatz, Claudia Wagner, Daniel Russell, Daricia Wilkinson,, David Piorkowski, April Yi Wang, Casey Dugan, Chacha Chen, Chenhao Tan, Dakuo Wang, Daniel Gruen, Elizabeth Daly, Finale Doshi-Velez, Gagan Bansal, Gloria Mark, Hal Daume III, Hariharan Subramonyam, Javier Antorán, Jennifer Wang, Jiao Sun, Jennifer Wortman Vaughan, Jian Zhao, Jina Suh, Jonathan Dodge, Justin D. Weisz, Kartik Talamadupula, Khai Truong, Klaus Mueller, Koustuv Saha, Kush R. Varshney, Mark Riedl, Marina Danilevsky, Matthew Davis, Markus Strohmaier, Mayank Agarwal, Michelle Zhou, Michael Muller, Michael Hind, Mingming Fan, Mohit Jain, Nikola Banovic, Öznur Alkan, Peter Pirolli, Prasanna Sattigeri, Pratyush Kumar, Praveen Chandar, Rachel Bellamy, Rajan Vaish, Ronny Luss, S. Shyam Sundar, Samuel Carton, Samir Passi, Sarah Miller, Shamsi T. Iqbal, Shreya Chowdhary, Shweta Narkar, Shwetak Patel, Snehal Prabhudesai, Soya Park, Stephanie Houde, Steven Moore, Su Lin Bloddget, Sumit Asthana, Susu Zhang, Thomas Erickson, Umang Bhatt, Upon Ehsan, Valerie Chen, Victoria Bellotti, Vivian Lai, Wai-Tat Fu, Werner Geyer, Xianyou Yang , Xun Huan, Yiming Zhang, Yunfeng Zhang, Yunyao Li, Zahra Ashktorab, Ziang Xiao